MINT 2026

Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT)

Proceedings of the Workshop

October 29, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-482-8

Introduction

Welcome to MINT, the first workshop on *Multimodal Interaction in Face-to-Face Dialogue*, co-located with EMNLP 2026 and taking place on October 29, 2026, in Budapest, Hungary.

This first edition brings together researchers from computational linguistics, natural language processing, computer vision, and cognitive science to explore multimodal verbal and non-verbal behaviours in face-to-face communication. The workshop covers models that integrate verbal and non-verbal cues; empirical studies of multimodal communication; datasets combining synchronised speech, video, and motion; methods for evaluating multimodal understanding and generation; and applications such as embodied conversational agents, social robots, and virtual avatars. With these themes, we want to connect the study of human communication with the development of AI systems.

We received 33 direct submissions and 2 ARR commitments, of which 18 were accepted as archival papers and four as non-archival papers. We thank all the authors for their contributions and the program committee members for dedicating their time and expertise to shaping the program. We are also excited to welcome our two invited speakers, Vera Demberg (Saarland University) and Judith Holler (Radboud University and Max Planck Institute for Psycholinguistics), and thank them for sharing their research with us.

Finally, we are grateful to the Multimodal Language Department (MLD) at the Max Planck Institute for Psycholinguistics for sponsoring this event. We look forward to a stimulating exchange of ideas and thank everyone for being part of the first MINT workshop.

The MINT organisers: Esam Ghaleb, Ece Takmaz, Diego Frassinelli, Bulat Khaertdinov, Raquel Fernández, Zerrin Yumak, Asli Ozyurek.

Sponsors

The MINT organizers are grateful for the support from the Multimodal Language Department (MLD) at Max Planck Institute for Psycholinguistics (www.mpi.nl).

Program Committee

Program Chairs

Esam Ghaleb, Max Planck Institute for Psycholinguistics
Ece Takmaz, Utrecht University
Diego Frassinelli, LMU Munich
Bulat Khaertdinov, Maastricht University
Raquel Fernández, University of Amsterdam
Zerrin Yumak, Utrecht University
Asli Ozyurek, Max Planck Institute for Psycholinguistics / Radboud University

Program Committee

lya Burenko, TU Dresden
Hendrik Buschmeier, Bielefeld University
Kerstin Fischer, University of Southern Denmark
Patrick Healey, Queen Mary University of London
Gustav Eje Henter, KTH Royal Institute of Technology
Hayley Hung, TU Delft
Hyewon Jang, Göteborg University
Ruben Janssens, Ghent University
Casey Kennington, Boise State University
Taras Kucherenko, National Library of Sweden
Lanmiao Liu, Max Planck Institute for Psycholinguistics
Rachel McDonnell, Trinity College Dublin
Chinmaya Mishra, Max Planck Institute for Psycholinguistics
Patrizia Paggio, University of Malta
Catherine Pelachaud, Sorbonne University
David Schlangen, University of Potsdam
Gabriel Skantze, KTH Royal Institute of Technology
Hugh Mee Wong, Utrecht University

Invited Speakers

Vera Demberg, Saarland University & Max Planck Institute for Informatics
Judith Holler, Radboud University & Max Planck Institute for Psycholinguistics

Table of Contents

Multi-modal, Multi-Speaker Dialogue Turns Modeling with Multi-modal LLMs
Lukas Hilgert and Jan Niehues 1

From Output Metrics to Cognitive Process: Disfluency and Lexical Entrainment for Evaluating Real-Time AI-Mediated Communication
Wanqing Psyche He and Susan Fussell 10

Cadence-Stratified Multimodal Perception for Real-Time Child–Robot Interaction
Marwan Mohammed, Javad Zarrin, Salma ElSayed and Vera Kempe 22

Multimodal Code-Mixed Hindi-English Dataset For Humor Recognition in Conversations
Abhishek Singh Bisht, Rakesh Kumar and Amrita Chaturvedi 33

Heads That Ask: Prosody of Questionhood in Whispering and Mutual (In)visibility of Speakers
Nasim Mahdinazhad Sardhaei, Marzena Zygis and Hamid Sharifzadeh 40

Exploring Multimodal Turn-Taking Cues in Face-to-Face Conversation using Voice Activity Projection
Willem Berner, Julio Cesar Cavalcanti, Kalle Åström and Gabriel Skantze 50

Automatic Annotation of Co-speech Gesture for Multimodal Communication Research – A Framework and a Case Study
Ilya Burenko, Peter Uhrig and Anna Wilson 59

Same Motions, Different Winners: The Gesture-Generation Leaderboard and Its Human Agreement Are Configuration-Tunable
Deep Gandhi, Ali Asaria and Tony Salomone 72

Behavioral Alignment and the Engagement for Non-Speaking Participants in Online Meetings
Trung Duc Anh Dang, Zhiqing Peng and Patrizia Paggio 78

Including Non-Manual Markers in Studies of American Sign Language Dialogue
Jillian Stebbins, Amber Hamilton, Emily Nover, Stephanie Patterson, Allison Fitch, Patrick Boudreault, Raja Kushalnagar and Cecilia O Alm 89

Do Motion Tokenizers for Co-Speech Gesture Generation Encode Gesture Semantics?
Varsha Suresh, Divij Jain, Jia Liu, M. Hamza Mughal and Vera Demberg 96

Looking for Objects and Looking for Help: Robot Gaze Elicits Human Cooperation
Anna Manaseryan, Thi Lan Anh Bui and Casey Kennington 105

Speaker gaze and disfluencies during storytelling toward a robot or a human: The role of partner type and temperamental shyness
Nils Frederik Tolksdorf and Adriana Hanulíková 117

FriendBench: Benchmarking Dyadic Familiarity Inference in Humans and Multimodal Large Language Models
Jeffrey M. Girard, Jason Zheng, Jacqueline Romero Vertino, Antony D'Avirro and Benjamin Peloquin 127

MentaLED: A Multimodal Conversational Agent for Mental Health Support Using Linguistic and Paralinguistic Speech Cues
Changhyun Kim, Dahye Kim, Eunseon Seong, Harim Lee, Sangyool Lee and Dong-Kyu Chae 144

Annotating and Recognizing the Palm Up Open Hand Gesture in Dutch
Fenna Berger, Alan Cienki and Lucia Donatelli 150

The Hidden Cost of Personality: How Conditioning Methods Disrupt Gesture Generation in Small Language Models
Tova Radhe and Alessio Galatolo 160

Gaze as Evidence for Common Grounding: A Cross-Corpus Analysis of MapTask and MUNDEX
Nan Li, Albert Gatt and Massimo Poesio 175

Multi-modal, Multi-Speaker Dialogue Turns Modeling with Multi-modal LLMs

Lukas Hilgert **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{lukas.hilgert, jan.niehues}@kit.edu

Abstract

State-of-the-art dialogue systems focus on two-person dialogue. A major challenge when extending this to multiple speakers is turn-taking – as it is nontrivial for more than two speakers. In this work, we investigate how this task can be formulated and how multi-modal LLMs (MLLMs) are able to perform it. First, we frame the task as multi-modal Next Speaker Prediction (mNSP) and create a multi-modal, multi-speaker NSP dataset from existing dialogue datasets. Second, we build dedicated systems for the mNSP task using MLLMs, and finally perform automatic evaluation, comparing them to rule-based baselines. We also conduct a human baseline study.

1 Introduction

In multi-speaker dialogues, each participant needs a clear cue for when it is appropriate to speak. Interrupting at an untimely moment can irritate other speakers and hinder the achievement of conversational goals. For dialogue systems, mastering this timing is essential. Speaking too frequently may annoy the user, whereas speaking too rarely can make the system seem unresponsive and defeat its purpose.

Previous work (Hilgert and Niehues, 2025) focused on predicting the next speaker from ground-truth text transcripts of dialogues with explicit speaker diarization. However, for humans, multi-modal clues are essential to correctly decide when to speak. In a dialogue, participants often provide visual cues while speaking such as gestures, hinting which other speaker they expect to react to their statement. Through audio cues, they can indicate when they begin or end their turn to speak (Schegloff et al., 1996; Ogden, 2013).

Recent success of multi-modal Large Language Models (MLLMs) with their ability to jointly model visual and audio information enable us to successfully combine the different input types. We frame the ability to utilize vision and audio perception for turn-taking as the multi-modal Next Speaker Prediction (mNSP) task (similar naming to Hilgert and Niehues (2025)).

In this work, we investigate whether and how well MLLMs can do the mNSP task, as follows:

- We construct a multi-modal training and benchmark dataset for mNSP from three existing dialogue datasets.
- We evaluate rule-based baselines and state-of-the-art MLLMs on the benchmark.
- We build dedicated mNSP models by fine-tuning MLLMs.
- We perform a human baseline study.

2 Multi-modal NSP Benchmark

We define the mNSP task like Hilgert and Niehues (2025) but extend it to multi-modal input: given a video clip together with accompanying audio of the last turns of a dialogue, a model should predict the next speaker based only on the information from the video and audio modality.

As a first step, we designed a benchmark based on the AMI meeting corpus (Carletta et al., 2005), the MULTISIMO multi-modal corpus (Koutsombogera and Vogel, 2018), and the MSDWild multi-modal speaker diarization dataset (Liu et al., 2022). AMI consists of recorded meetings of four people. It provides video and audio of the meetings together with a transcript that also shows when each speaker spoke. MULTISIMO contains discussions between two players and a facilitator about a game-like task / quiz. It provides the same level of detail as AMI but is however a lot smaller by total recording duration (Table 1). The MSDWild dataset is intended for speaker diarization. It is sourced from public videos ("YouTube Vlogs") with two to ten speakers and is manually labeled. In contrast to the other two datasets, the language spoken is not only English but also Chinese, Thai, Korean, Japanese,

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 1–9
October 29, 2026 ©2026 Association for Computational Linguistics

	AMI	MULTISIMO	MSDWild
Duration	100h	3h	80h
Dialogues	162	17	1015
Turns	74377	3807	18747
train	80.96	0.00	79.31
dev	10.82	44.86	9.35
test	8.22	55.14	11.35

Table 1: Total runtime of the original datasets, the total number of dialogues and mNSP task items ("turns") after our filtering, and the percentage of turns per split for each dataset (see subsection C.1).

German, Portuguese, and Arabic. Also, it provides no ground-truth transcript of the audio spoken in the videos.

We split AMI according to the split provided by the dataset authors. As the MULTISIMO dataset has no official split and is fairly small, we create our own 1:1 split (dev / test) that is for evaluation only (Table 1). Most of the videos in MSDWild involve only two speakers; therefore, we ignore the split suggested by the dataset authors and create a new one (train / dev / test) that contains only dialogues with more than three speakers. We list more details in subsection C.1.

To eliminate ambiguous turns, we discard any turn that overlaps with the subsequent turn, while retaining both utterances in the context. The initial turn of each dialogue is removed because, without prior context, an informed decision cannot be made. Speaker identifiers from the dataset (e.g., "MTD011UID") are replaced with pseudonyms of the form "speaker <number>".

3 Multi-modal NSP with LLMs

To model the mNSP task as an MLLM task, we prompt the MLLMs to predict the next speaker by utilizing the information we provide in the prompt (Appendix A); an instruction for the task and the four most recent turns of the current dialogue as context. We use only the last four as this was shown to perform not much worse than eight or 16 turns in a text-only setting (Hilgert and Niehues, 2025) and we observe no high performance difference in our setting (Table 5). We source the beginnings and ends of each turn from the timestamps provided by the datasets.

Additionally, we have to insert speaker-identifying information so that the model's output can be evaluated easily. To inform the model how we assigned identifiers to the speakers, we insert the name (speaker <number>) and an image (screenshot from the video) of each speaker in the prompt.

We fine-tune the models supervised on data from the training set from AMI and MSDWild where the expected output text is the next speaker identifier based on the aforementioned prompt.

4 Experiments

We fine-tune MLLMs and compare them to simple rule-based and model-based methods.

4.1 Setup

We share our full training setup here and further details in Appendix B.

Models We use and fine-tune the 3B parameter model of the Qwen2.5-Omni family (Xu et al., 2025a) of open-weight MLLMs because they offer native audio, image, and video input. We compare the performance to zero-shot Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) and fine-tuned LLaVA-Next-Video-7B-DPO (Zhang et al., 2024).

Training Data As training data, we use the training split from our composite dataset. We only train on those samples where the duration of the video clip of four turns takes less than 30 seconds to avoid out-of-memory issues during training. This requirement holds for most of the data points (98.4%), so we assume that excluding those data points will not substantially degrade performance.

Ablation Studies We investigate how using only one of the two datasets with training sets (AMI and MSDWild) influences performance (Table 5). In similar experiments, we train on more than four turns as context. Additionally, we fine-tune one adapter to first do speaker diarization by listing all involved speakers in the provided clip in order of occurrence and to then predict the next speaker (**two-step prompt**). This data is available for MSDWild (even though the transcript is missing) and for AMI.

To show the benefit of direct multi-modal input, we check our results against a **text-only** baseline where an MLLM transcribes the conversation (with speaker diarization) zero-shot and based on that an MLLM fine-tuned on transcripts predicts the next speaker. To transcribe the dialogue, we provide the model with the same information as the direct multi-modal setup (video with audio of the conversation and an image of each speaker). We also

run this experiment with the ground-truth dialogue transcript. These, however, are only available in the AMI and MULTISIMO dataset.

Metrics We select multiple metrics to evaluate performance on mNSP to display different aspects of the task and the datasets. As our primary evaluation metric, we use next-speaker accuracy because it directly reflects the available data and task. Since the distribution of turns per speaker is imbalanced (Table 4), we also report the weighted F_1 score. Finally, because a few baselines are deterministic (yet high performing), we include the Matthews correlation coefficient (MCC) to verify that our models outperform random chance. Each score is calculated per dialogue and then aggregated with weighted (dialogue length) average.

4.2 Rule-based Baselines

We report evaluation results for four simple, rule-based baselines. Two of them model opposite dialogue-flow patterns, a third combines both approaches, and a fourth always picks the same speaker. The baselines are:

Baseline	Definition
Always switch	Random speaker, but forced to switch after every turn
Never switch	Predict the same speaker as in the previous turn
Random	Completely random speaker, allowing repeats
Speaker 0	Always select the speaker who spoke first in the dialogue

Both the **Always switch** and **Never switch** baselines require knowledge of the true last speaker (ground-truth information). Consequently, we also provide a second **Never switch** baseline (as its performance is in general higher than **Always switch** and therefore more interesting) that uses the last-speaker prediction generated by our two-step prompt setup instead of the true label.

4.3 Human Baseline Study Setup

We ran a preliminary human baseline study with six participants on one dialog (168 turns) from the AMI dataset's test split. We divided the dialogue into three parts (1, 2, 3) to limit the workload for each labeler. We let each part be labeled by two labelers and average their scores. We present more details in Appendix D.

The labelers were presented with the same prompt and information (video + images) for their decision as the MLLM. We compare the human baseline against the best system when evaluating on the full AMI test split (Qwen2.5-Omni-3B fine-tuned for two-step prompt).

4.4 Results

We analyze the performance of the baselines and our fine-tuned models.

Rule-based Baselines The **Never switch** baseline (both ground truth and prediction variant) performs the best with both accuracy and F_1 score. This resonates with a characteristic of the AMI and partially the MSDWild dataset that speakers often have multiple turns after another. Consequently, the method where the speaker always switches performs very poorly for AMI. We observe this also with MCC scores as they are only in the (higher) positive range for "Never". The method with the least assumptions that picks the next speaker completely randomly ("Random") performs roughly on the expected random chance of 25% for a four-speaker dialogue (AMI) and 33% for three speakers (MULTISIMO).

A general takeaway is that even though the dataset is imbalanced (lower speaker ID $\rightarrow$ more turns, Table 4), the accuracy and F_1 scores have close values for most (rule-based) experiments (Table 2). An exception is "Speaker 0". As this baseline always predicts only one specific speaker, the accuracy is high (as speaker 0 has the highest share of turns) but the F_1 score is significantly lower as no true positives for the other speaker exist.

MLLMs The models show an expected difference; the bigger model performs better across almost all metrics and datasets in the zero-shot setting (Table 2).

The **Never switch** baseline has the advantage of ground truth information about the last speaker. This becomes clear when we rely on the last speaker being predicted by an MLLM (**Never switch (prediction)**) and scores drop below our fine-tuned MLLMs on all datasets and metrics (except for F_1 on MULTISIMO). We also see that in general, the models perform better for turns where no speaker switch occurs (Table 5, speaker switch vs. same speaker). We conclude that, rather than merely prediction the most common or previous speaker, fine-tuned MLLMs capture richer turn-taking patterns.

Subset	AMI			MULTISIMO			MSDWild		
Metric	Acc	F_1	MCC	Acc	F_1	MCC	Acc	F_1	MCC
Rule-based Baselines									
Always switch	15.06	15.63	-11.86	34.23	34.47	2.08	25.17	26.24	-1.50
Random	24.83	26.02	-0.12	32.08	**32.38**	-1.96	29.50	31.15	0.28
Never switch (GT)	55.02	55.29	36.56	31.25	31.44	-4.61	39.42	39.38	4.02
Never switch (prediction)	35.78	35.75	11.91	33.11	31.57	-0.53	32.60	32.37	2.53
Speaker 0	30.49	15.91	0.00	35.97	19.48	0.00	37.85	22.98	0.00
Zero-shot MLLM									
Qwen2.5-Omni-3B	19.70	13.41	0.38	28.20	22.06	0.70	22.99	18.99	-0.37
Qwen3-Omni-30B	30.62	27.59	5.48	30.20	24.28	-2.03	31.64	28.16	0.00
Ground truth text transcript as input for MLLM (Qwen2.5-Omni-3B), no multi-modal information									
Zero-shot	12.42	13.80	-9.55	21.92	25.78	3.67	-	-	-
Fine-tuned on AMI	59.16	59.12	41.81	39.21	38.46	6.14	-	-	-
Text-only MLLM (first transcribe zero-shot, then predict)									
Qwen2.5-Omni-3B	29.07	23.05	-0.62	34.11	29.96	2.42	33.76	28.68	1.06
Fine-tuned MLLM									
Qwen2.5-Omni-3B	48.55	45.48	26.61	**37.68**	27.77	**7.10**	36.20	31.06	1.23
LLaVA-Next-Video-7B	30.13	16.06	0.57	35.02	24.57	-0.81	36.34	29.07	-0.40
Fine-tuned MLLM, two-step prompt									
Qwen2.5-Omni-3B	**49.26**	**46.84**	**26.68**	36.15	27.42	1.21	**38.27**	**34.21**	**4.72**

Table 2: mNSP on the test set. We compare simple rule-based baselines with the MLLMs fine-tuned on different task setups. All values are scaled by 100 for better readability. Scores from experiments with direct ground-truth (GT) model input are grayed out.

Ablation Studies Running the next speaker prediction task after performing speaker diarization explicitly (**two-step prompt**) shows small but consistent gains over the direct prompt for AMI and MSDWild (Table 2).

When we let the model first transcribe the conversation and then predict the next speaker solely from the text, performance lags behind both the direct and the two-step prompt for almost all cases (Table 2). With the ground-truth transcripts and a fine-tuned MLLM, however, we exceed all other methods (except for the MCC score for MULTISIMO) by roughly 10 to 15 percentage points on AMI and two to six on MULTISIMO. As this method still is not near 100%, we think it shows again the task's ambiguity and that transcribing the dialogue (either implicitly or explicitly) is only one part of solving the mNSP task.

With LLaVA-Next-Video, we see that our fine-tuning also allows other MLLMs to perform the mNSP task; however, it falls short of Qwen2.5-Omni likely because of the older base model and the missing audio modality.

Training and inference on more than four turns increases performance consistently but only slightly (Table 5), albeit with longer training and inference times. In similar experiments, we saw that training on both AMI and MSDWild leads to more robust predictions (Table 5).

In general, fine-tuning MLLMs on dialogue data—by aligning video and audio with the next speaker—enhances mNSP capabilities above rule-based baselines when ground-truth is unavailable. Moreover, this improvement is non-trivial, as implied by the MCC scores.

4.5 Human Baseline Study Results

Our human baseline study shows that the average human performance is close to our system's average performance. They also both have drastically reduced mNSP scores on the second part of the dialogue compared to other parts. For the first part, our system underperforms the human labelers while for the third part, our system achieves higher scores on all metrics. So even though humans are able to correctly and reliably pick the next speaker in a multi-speaker dialogue setting, they do not outperform our system by a wide margin in this study. We think this shows both the strength of our approach and highlights the ambiguity of the mNSP task.

Part of AMI subset	Metric	Average from two labelers	Fine-tuned MLLM, two-step prompt
1	Acc	64.29	53.57
	F1	63.68	49.44
	MCC	35.30	-2.97
2	Acc	36.61	33.93
	F1	32.88	32.46
	MCC	11.15	6.47
3	Acc	40.18	51.79
	F1	40.08	44.70
	MCC	1.61	8.51
Average	Acc	47.02	46.43
	F1	45.54	42.40
	MCC	16.02	8.55

Table 3: Results from human baseline study on a subset of the AMI test set; numbers from Qwen2.5-Omni-3B fine-tuned for two-step prompt as comparison.

5 Related Work

Previous research on dialogue turns differs from our approach, as we assume a multi-speaker dialogue setting with an MLLM as the predictor. More theoretical work focuses on the fundamental rules of turn-taking (Sacks et al., 1974) and the influence of audio (Schegloff et al., 1996; Dhillon et al., 2004; Ogden, 2013) and visual clues (Kendon, 1967; Kendrick et al., 2023).

Transition Relevance Places Methods for turn-taking use LLMs to predict transition-relevant places within a stream of words. Transition-relevant places are points in a dialogue where a turn-shift can happen. Ekstedt and Skantze (2020) fine-tuned GPT-2 to predict these spots in written and spoken dialogues. Later work (Umair et al., 2024) experimented with more recent LLMs. This research focused on only two simultaneous speakers. Newer work extends this to multiple speakers but only considers the text domain (Hilgert and Niehues, 2025).

Audio / Visual Cues Multi-modal approaches for next-speaker prediction (NSP) exploit visual cues such as gaze and hand gestures (Ishii et al., 2016; Malik et al., 2020). This research incorporates gaze-transition patterns and the timing structure of eye contact (Ishii et al., 2016) or head movements (Ishii et al., 2015). Malik et al. (2020) further employed focus-of-attention features, among others, to train classic machine-learning classifiers for NSP.

Other systems base turn-taking prediction on voice-activity projection (VAP), forecasting future voice activity from the current audio signal (Inoue et al., 2024a,b). These methods have been extended to handle multiple speakers (Elmers et al., 2025). A recent end-to-end framework combines a VAP model with a turn-taking model for human-robot conversation (Skantze and Irfan, 2025).

6 Conclusion

Our research goal is to investigate the ability of fine-tuned multi-modal LLMs to predict the next speaker in a multi-speaker dialogue setting using multi-modal information (mNSP task). We create a multi-modal benchmark from the existing dialogue dataset AMI, MULTISIMO, and MSDWild. The experiments on our benchmark show that small fine-tuned MLLMs like Qwen2.5-Omni-3B can outperform rule-based baselines and pipeline approaches (transcribe, then predict) on accuracy, F_1, and MCC score – as long as they do not rely on ground truth information.

In future work, we will investigate the ambiguity of the task (should vs. could speak) and how humans handle it.

Limitations

We do not predict the upcoming speaker on a token-by-token or word-by-word basis; instead, a prediction occurs after a complete utterance (as defined by datasets timestamps) has finished. Consequently, the system expects whole utterances as input, not a continuous audio-video stream. This workflow can be realized by feeding our system with the output of a multi-speaker VAP model, e.g., the one proposed by Elmers et al. (2025).

Predicting only one speaker as the next possible one may not consider the ambiguity of the task. Multiple speakers could be eligible if, for example, one speaker asks an open question to two other speakers where both are supposed to answer. We limit our work to predicting only one speaker at a time because it is more straightforward both for evaluation and training given the existing datasets. We also assume that predicting the next speaker for thousands of data points helps average out the ambiguity and still gives a good view on the (m)NSP capabilities of (M)LLMs.

Acknowledgments

The research leading to these results was supported by the European Union's Horizon Europe programme grant agreement No. 101213369 (DVPS) and the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried M. Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2981–2990. Association for Computational Linguistics.

Mikey Elmers, Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2025. Triadic Multi-party Voice Activity Projection for Turn-taking in Spoken Dialogue Systems. In *Interspeech 2025*, pages 3015–3019.

Lukas Hilgert and Jan Niehues. 2025. Next speaker prediction for multi-speaker dialogue with large language models. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 60–71, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11873–11883. ELRA and ICCL.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and continuous turn-taking prediction using voice activity projection. *CoRR*, abs/2401.04868.

Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 2319–2323. IEEE.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6(1):4:1–4:31.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473.

Maria Koutsombogera and Carl Vogel. 2018. Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tao Liu, Shuai Fan, Xu Xiang, Hongbo Song, Shaoxiong Lin, Jiaqi Sun, Tianyuan Han, Siyuan Chen, Binwei Yao, Sen Liu, Yifei Wu, Yanmin Qian, and Kai Yu. 2022. MSDWild: Multi-modal Speaker Diarization Dataset in the Wild. In *Proc. Interspeech 2022*, pages 1476–1480.

Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who speaks next? turn change and next speaker prediction in multimodal multiparty interaction. In *32nd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2020, Baltimore, MD, USA, November 9-11, 2020*, pages 349–354. IEEE.

Richard Ogden. 2013. Clicks and percussives in english conversation. *Journal of the International Phonetic Association*, 43(3):299–320.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Emanuel A Schegloff and 1 others. 1996. Turn organization: One intersection of grammar and interaction. *Studies in interactional sociolinguistics*, 13:52–133.

Gabriel Skantze and Bahar Irfan. 2025. Applying general turn-taking models to conversational human-robot interaction. In *20th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2025, Melbourne, Australia, March 4-6, 2025*, pages 859–868. IEEE.

Muhammad Umair, Vasanth Sarathy, and Jan Peter de Ruiter. 2024. Large language models know what to say but not when to speak. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 15503–15514. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *CoRR*, abs/2509.17765.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

A Prompt

Here, we present the prompt we employed for the mNSP task:

```
Your task is to predict the next speaker given the full conversation history. Do not provide any explanation. Do not complete the conversation.

This is a video clip of the conversation:
<video>
This is an image of speaker 0:
<image>
This is an image of speaker 1:
<image>
This is an image of speaker 2:
<image>
This is an image of speaker 3:
<image>

Predict the next speaker by outputting the name and only the name of the next speaker. Carefully consider the motives of the participating speakers in the conversation. Do not provide any explanation. Do not complete the conversation.
```

B Inference and Training Details

- Models
  - Qwen2.5-Omni-3[1]
  - Qwen3-Omni-30B-A3B-Instruct[2]
  - LLaVA-NeXT-Video-7B-DPO[3]
- Inference
  - Hugging Face Transformers library (Wolf et al., 2019) for loading and running the models.[4]
  - Temperature: 0.0 (no sampling)
- Training
  - LlamaFactory (Zheng et al., 2024)[5] with FlashAttention 2 (Dao, 2024).
  - In general, we followed the example script for Qwen2.5-Omni.[6]
  - Supervised LoRA (target all layers) fine-tuning (Hu et al., 2022) with rank $r = 8$.
  - adamw_torch optimizer
  - Learning rate 1.0×10^{-4}, cosine LR scheduler, 0.1 warmup ratio.

[1] https://huggingface.co/Qwen/Qwen2.5-Omni-3B
[2] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
[3] https://huggingface.co/lmms-lab/LLaVA-NeXT-Video-7B-DPO
[4] https://github.com/huggingface/transformers
[5] https://github.com/hiyouga/LLaMA-Factory
[6] https://github.com/hiyouga/LlamaFactory/pull/7537

  - One training run per experiment.

- Hardware equipment: Up to three NVIDIA RTX 6000 Ada Generation GPUs or one NVIDIA RTX PRO 6000 Blackwell at the same time on a small-scale GPU server cluster.

C Dataset Details

We provide more details on our preprocessing of the datasets here and statistics of the resulting filtered datasets.

C.1 Preprocessing

- Split
  - We split the dataset either using the training/validation/test split provided by the original authors (AMI) or a custom split (MULTISIMO, MSDWild).
  - We assign entire dialogues to splits without dividing any single dialogue across them.
  - We define a target ratio (e.g., 80% / 10% / 10% for MSDWild), shuffle the dialogues randomly using Python's `random` module ($seed = 42$), and then allocate them according to the desired ratios.

- Filtering
  - Conversations from the AMI dataset involving fewer or more than four speakers (9 out of 171 dialogues) are excluded to simplify performance comparison across dialogues for AMI.
  - For MSDWild, we only include those dialogues with more than two speakers.
  - In the MULTISIMO dataset, all dialogues have exactly three speakers.
  - Turns that exhibit mismatches between audio and video files are omitted.

C.2 Speaker Contribution

Our analysis (Table 4) shows that the speaker that appears earlier in the conversation has a higher share of turns per dialogue (if all dialogues have the same number of speakers (AMI, MULTISIMO)). For MSDWild, the pattern is similar but also impacted by the fact that most dialogues have only three speakers (so e.g., speaker 3 will appear less often).

Dataset	AMI	MULTISIMO	MSDWild
Speaker 0	30.49	35.97	37.85
Speaker 1	25.94	30.40	23.27
Speaker 2	24.80	33.64	29.62
Speaker 3	18.77	-	7.90
Speaker 4	-	-	0.99
Speaker 5	-	-	0.33
Speaker 6	-	-	0.00
Speaker 7	-	-	0.05

Table 4: Contributed turns (in percentage) from each speaker across all dialogues (test set).

D Annotation Acquisition for Human Baseline Study

We asked colleagues working in the field of NLP to fill out a web-based form for our user study to acquire a human baseline. We created a website for labeling each video clip with the next speaker using Gradio (Abid et al., 2019).[7] The participation was not mandatory, and we offered no compensation. We informed the participants that the data created by them during this user study will be incorporated into a scientific publication.

E AI Assistance

AI assistance for writing and coding was only used in a way that it did not impact the findings presented in this paper.

F Additional results

We share additional results in this section (Table 5) that are not central to our findings.

[7] https://gradio.app/

Subset	AMI			MULTISIMO			MSDWild		
Metric	Acc	F_1	MCC	Acc	F_1	MCC	Acc	F_1	MCC
Fine-tuned MLLM (Qwen2.5-Omni-3B), AMI + MSDWild, vary context									
4 turns	48.55	45.48	26.61	37.68	27.77	7.10	36.20	31.06	1.23
8 turns	48.93	46.55	28.38	37.68	31.39	9.05	37.56	33.10	1.36
Rule-based Baselines, evaluate different turns									
Never switch (GT)	55.02	55.29	36.56	31.25	31.44	-4.61	39.42	39.38	4.02
Speaker switch	0.07	0.07	-35.84	0.26	0.30	-50.02	1.49	1.45	-46.32
Same speaker	99.85	99.85	99.78	99.88	99.88	99.81	96.63	96.90	88.00
Fine-tuned MLLM (Qwen2.5-Omni-3B), evaluate different turns									
AMI + MSDWild	48.55	45.48	26.61	37.68	27.77	7.10	36.20	31.06	1.23
Speaker switch	35.88	31.83	10.01	37.52	27.65	9.01	33.25	27.63	1.41
Same speaker	58.89	56.56	40.09	39.82	30.52	12.30	42.33	40.01	3.23
Fine-tuned MLLM (Qwen2.5-Omni-3B), different training mix									
AMI + MSDWild	48.55	45.48	26.61	37.68	27.77	7.10	36.20	31.06	1.23
AMI	50.68	47.96	29.17	32.35	23.99	4.34	38.36	31.16	3.12
MSDWild	34.15	32.91	10.71	39.07	30.32	6.53	38.93	33.03	3.36

Table 5: mNSP on the test set. We compare simple rule-based baselines with the MLLMs fine-tuned on different task setups. All values are scaled by 100 for better readability.

From Output Metrics to Cognitive Process: Disfluency and Lexical Entrainment for Evaluating Real-Time AI-Mediated Communication

Wanqing Psyche He and Susan R. Fussell
Department of Information Science, Cornell University
Ithaca, NY, USA
{wh385,sfussell}@cornell.edu

Abstract

Real-time communication mediated by artificial intelligence (AI) can produce an accurate output while changing how a speaker plans an utterance and what a listener receives. Output measures, such as word error rate, machine translation scores, and task accuracy, leave these process effects unobserved. We propose a dual-axis framework that aligns disfluency measures with suggestion correspondence, including lexical entrainment and semantic similarity. Disfluency type, rate, timing, and duration characterize observable production difficulty. Correspondence measures characterize whether suggestion content appears in the response. Observing both axes within the same episode distinguishes behavioral configurations that either family alone leaves ambiguous and supports explicit hypotheses about planning, uptake, and resistance. An illustrative corpus demonstrates event-detection feasibility and shows that semantic correspondence varies with coded incorporation and exceeds a comparison matched on presentation format. These checks leave cognitive interpretations open to experimental validation. We also derive a prediction concerning listeners: removing natural disfluencies may remove cues that listeners use to anticipate difficult or novel material. The resulting framework and reporting protocol extend standard evaluation to the human processes through which real-time AI-mediated communication succeeds or fails.

1 Introduction

Natural language processing (NLP) evaluation of spoken dialogue systems has long asked whether a system produced an accurate or useful output. Word error rate (WER) evaluates automatic speech recognition (ASR), Bilingual Evaluation Understudy (BLEU) and chrF evaluate machine translation (MT), and task completion evaluates goal-oriented dialogue (Papineni et al., 2002; Post, 2018). These measures have driven progress in speech recognition, spoken language understanding, and end-to-end (E2E) dialogue modeling (Shriberg, 1994; Jurafsky and Martin, 2024). Their unit of analysis becomes incomplete when a system intervenes during communication itself. Real-time AI-mediated communication (AI-MC) now supplies suggestions, lexical clarifications, and spoken completions while a person is planning and producing speech (Duan et al., 2021; He and Fussell, 2026; Qin et al., 2025; Fang et al., 2025; Xia et al., 2023). Two systems can consequently receive the same output score while differing in planning disruption, response reuse, or preservation of cues that support a listener. Evaluation that overlooks these differences can favor a system that improves the observable product while degrading the interaction through which it was produced.

Speech provides evidence about these otherwise hidden process differences. Speakers produce approximately 2–3 words per second while coordinating lexical selection, phonological encoding, prosodic planning, and articulation (Levelt, 1989). Filled pauses, repetitions, and restarts occur in roughly 6–10% of spoken words in naturalistic conversation (Bortfeld et al., 2001; Fox Tree, 1995), and their distributions vary with planning, retrieval, monitoring, and interaction management (Goldman-Eisler, 1968; Shriberg, 1994; Clark and Tree, 2002; Corley and Stewart, 2008). An event still has several possible sources because its interpretation depends on the speaker, task, conversational position, and linguistic context. Its type and timing can localize observable difficulty, while correspondence between the AI suggestion and the response establishes whether related content appeared. The two forms of evidence answer complementary questions about the same episode.

We therefore advance a framework with two observable axes: *speech measures* and *suggestion correspondence*. Lexical entrainment provides the established word-choice measure within the second

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 10–21
October 29, 2026 ©2026 Association for Computational Linguistics

axis, while semantic and human-coded measures capture correspondence beyond shared wording. Aligning the axes yields four behavioral profiles that organize competing explanations without assigning a unique cognitive state. A reporting protocol then specifies the anchors, comparisons, and missing-data distinctions needed to test those explanations. The same process perspective produces an additional prediction for listeners, who may lose useful anticipatory cues when AI systems smooth disfluent speech (Arnold et al., 2003; Fox Tree, 2001; Corley et al., 2007). Our contribution is a way to formulate and evaluate these process questions alongside output quality. We begin by establishing what speech-production theory permits the measures to mean.

2 Theoretical Basis for Process Measures

Using disfluency as a process measure requires a theory that constrains its possible sources. Levelt's 1989 model organizes speech production into three incrementally operating stages: the ***conceptualizer*** generates a preverbal communicative intent, the ***formulator*** translates this intent into a linguistic plan, and the ***articulator*** executes the motor program as speech (Indefrey and Levelt, 2004; Levelt et al., 1999). A ***perceptual loop monitor*** compares inner speech and auditory feedback with intent, and detected mismatches can interrupt an ongoing plan (Roelofs, 2020; Nozari et al., 2011; Hartsuiker and Kolk, 2001).

Predictive processing accounts add that speakers and listeners anticipate utterances at multiple linguistic levels (Pickering and Garrod, 2013; Hagoort, 2019). Together, these accounts explain why external content arriving during production may affect planning, retrieval, monitoring, or management of the interaction. They also set an inferential boundary: event type and timing narrow candidate processes without identifying a unique stage.

2.1 Relation to NLP Research on Disfluency

NLP has usually treated disfluency as material to detect and remove. Work on the Switchboard corpus established taxonomies of interruption points, reparanda, and repairs (Shriberg, 1994), and subsequent sequence labelers, pretrained transformers, and E2E ASR rescorers filtered these events before parsing or MT (Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021). This approach supports downstream processing while discarding evidence about how that output was produced.

A separate literature shows that AI can change human language. Smart-reply and predictive-text systems narrow lexical diversity and shift register in text-based communication (Arnold et al., 2020; Mieczkowski et al., 2021; Hohenstein et al., 2023; Yakura et al., 2024), while work on conversational entrainment measures lexical and syntactic convergence between people (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). Real-time systems now connect these previously separate concerns: E2E models such as Moshi (Défossez et al., 2024) and SpeechGPT (Zhang et al., 2023), together with incremental turn-taking systems (Skantze, 2021), supply content during ongoing production. The resulting evaluation problem concerns both the system contribution and the human process of responding to it. This paper uses disfluency to preserve process evidence and suggestion correspondence to determine how that evidence relates to system content.

2.2 Disfluency Types and Production Processes

Disfluency categories become informative through their different distributions across production and interactional contexts. ***Silent pauses*** at clause and phrase boundaries are associated with conceptual planning and with the complexity or novelty of upcoming material (Goldman-Eisler, 1968; Butterworth, 1980). ***Filled pauses*** ("uh/um") occur during conceptual planning and lexical retrieval and can signal an upcoming delay to listeners (Clark and Tree, 2002).

Other categories provide stronger evidence about managing or revising an utterance. ***Repetitions*** can maintain the conversational floor while formulation continues (Schegloff et al., 1977; Engelhardt et al., 2010). ***Self-repairs*** provide overt evidence that a speaker halted and revised speech after monitoring (Hartsuiker, 2006; Levelt, 1983), and their onset can follow an error within hundreds of milliseconds (Hartsuiker and Kolk, 2001; Postma, 2000). Their functions overlap with effects of task difficulty, language background, turn management, and the surrounding utterance. Inference therefore depends on contrasts among types and positions. For example, a shift from fillers toward repairs is consistent with increased monitoring and correction, while remaining compatible with other contextual explanations.

2.3 How AI Mediation Perturbs Production

These associations between disfluency type and cognitive process make the production model experimentally useful. **Lexical clarifications** can resolve unfamiliar terms during comprehension and change the information available for response planning (Duan et al., 2021; He and Fussell, 2026). **Content suggestions**, such as topic prompts and factual completions, may change conceptual planning (Arnold et al., 2016; He and Fussell, 2026; Qin et al., 2025). **Lexical and stylistic suggestions**, such as word recommendations and register adjustments, supply candidate words or syntactic frames during formulation (Arnold et al., 2020; Hohenstein et al., 2023). **Timing cues**, such as turn-taking prompts, may change turn initiation and floor management (Skantze, 2021). These mappings specify predictions while allowing one intervention to affect several components.

Evaluating external content while preparing speech can place simultaneous demands on working memory (Oomen and Postma, 2001; Engelhardt et al., 2013). A nearby disfluency is consequently consistent with evaluation of the intervention and with unrelated production difficulty. This ambiguity separates *cognitive fluency*, the efficiency of planning and retrieval, from *utterance fluency*, observable speech production (Segalowitz, 2010). Lexical evidence, matched comparisons, and converging measures are required to connect an observable event to the intervention.

These mechanisms therefore motivate directional hypotheses. Relevant content available before a complex response may reduce pauses associated with planning. Content that arrives during formulation or conflicts with an emerging plan may increase repairs and repetitions. Manipulating production demand should chiefly alter speech event timing and frequency, whereas manipulating suggestion usefulness or attractiveness should chiefly alter suggestion correspondence. Crossing these manipulations tests whether the two measures respond selectively and predicts changes in the four profiles introduced next.

3 A Two-Axis Evaluation Framework

A useful framework must translate the theoretical distinction into observations that can be compared across systems. We organize those observations along two axes: speech measures identify when and how production difficulty becomes observable, and suggestion correspondence identifies whether system content appears in the response. Low disfluency can accompany smooth uptake, no consultation, silence, or withdrawal, and high correspondence can reflect considered use, automatic reuse, or content the speaker would have produced independently. Aligning the axes within the same episode reduces these complementary ambiguities and defines the evidence needed to distinguish the remaining explanations (Jakesch et al., 2023; Hohenstein et al., 2023).

Comparison also requires a common protocol. Studies should report anchors, windows, counting rules, baselines, and extraction procedures while adapting their values to the task and language. Unmediated baselines within each speaker are preferred when available. Studies without that condition should use matched windows or control trials and model variation among speakers. Raw disfluency values are unsuitable for unadjusted comparisons because individual rates vary substantially (Bortfeld et al., 2001).

3.1 Speech Measures

The speech axis begins with event identity and progressively adds duration, position, and change over time. Each layer restricts the plausible explanation of an event while preserving uncertainty arising from the speaker, task, and conversational context.

Disfluency rate by type. Report counts and rates (per 100 words, per minute, or per elapsed response time) of filled pauses, silent pauses, repetitions, repairs, and false starts. Stating the denominator makes systems comparable, and retaining categories preserves differences among their likely functions.

Disfluency duration. Report the temporal extent of each event because equal counts can conceal different amounts of interruption. Longer pauses and repair phases can accompany greater planning or reprocessing demands, with interpretation determined by event position and context (Hartsuiker, 2006).

Positional distribution. Locate event onsets relative to AI appearance, question completion, response onset, wording related to the suggestion, and response end. A filled pause after appearance is consistent with planning or lexical retrieval difficulty, whereas a repair after wording related to the suggestion provides evidence of revision. Posi-

tion connects an event to a candidate intervention period and operationalizes the *speech event timing* axis in §3.3. Alignment windows should be specified before analysis and tested for sensitivity to plausible boundary changes.

Disfluency trajectory. Report change across a conversation to distinguish initial disruption from adaptation or accumulated difficulty. Record silence, abandoned turns, and missing speech separately because each can otherwise appear as a low event rate.

3.2 Content Correspondence Measures

Once a speech event has been localized, the second axis asks how the response relates to the intervention. We distinguish surface lexical overlap, semantic similarity, and incorporation coded by people because each supports a different inference. *Suggestion correspondence* names the general axis, and *lexical entrainment* refers specifically to convergence in word choice across turns, following the dialogue literature (Nenkova et al., 2008; Niederhoffer and Pennebaker, 2002; Reitter, 2014).

Moving-average type-token ratio (MATTR). This measure summarizes lexical diversity across a session and uses a moving window to reduce the length sensitivity of raw type-token ratio (Covington and McFall, 2010). A decline can accompany repeated use of vocabulary suggested by AI (Hohenstein et al., 2023). It therefore detects change in expressive range across time. Attribution to an individual suggestion requires the correspondence measures that follow.

Suggestion correspondence. Compare the participant's response text with the displayed suggestion. Use surface measures such as n-gram overlap to assess wording reuse and sentence-embedding cosine similarity or BERTScore to assess semantic correspondence (Reimers and Gurevych, 2019; Zhang et al., 2020). Human coding can establish whether recognizable suggestion content appears. Reporting them separately prevents semantic similarity, lexical reuse, and incorporation from being treated as interchangeable. Together they operationalize response content while leaving endorsement and reliance to additional evidence.

Connotation and register shift. Compare the speaker's affective, formal, or stylistic profile with the AI output using formality classifiers (Pavlick and Tetreault, 2016), function-word style matching (Niederhoffer and Pennebaker, 2002), or affective representations. Persistence into later unmediated turns provides a stronger adaptation hypothesis than immediate reuse. Topic, audience, and interactional goals remain competing explanations that matched comparisons must address.

3.3 Combining Speech and Content Measures

The two axes become informative when measured for the same episode. Suggestion correspondence establishes the relation between response and system content, whereas speech event timing locates observable production difficulty around the content's availability and possible use. Crossing them yields four behavioral profiles (Table 1). The profiles convert two ambiguous signals into a structured set of competing explanations and specify which additional evidence could separate them.

Table 2 draws on a secondary analysis of the published XPLAIN Wizard-of-Oz study (He and Fussell, 2026). In that study, speakers using English as a second language completed a standardized collaborative planning task. The interface provided private, clickable lexical clarifications and presented predefined idea or sentence suggestions at scripted points. The source paper reports the system design and survey and interview findings. Here, timestamped recordings and response annotations serve only to demonstrate event alignment and correspondence measurement. Trained annotators coded disfluency events against intervention timestamps and coded suggestion incorporation from separate records. The excerpts illustrate measurement decisions and carry no validated profile labels.

Reading the excerpts descriptively, Row 1 shows close phrase correspondence in the reconstructed materials and a brief filler at retrieval onset. Row 2 contains a filler at clause onset outside the prespecified analysis window and uses a different lemma. Row 3 contains a pre-uptake *um*, the phonological *pack-/pack-aged/packable* repair, and content uptake of the AI's term "packable." Row 4 contains a silent pause and filler followed by lexical substitution within the AI's syntactic frame. These observations illustrate the component measures and remain compatible with multiple cognitive accounts.

The examples also show why a profile starts an analysis. Frequent Profile A observations could arise from smooth fit, uncritical reuse, or missed event detection, whereas frequent Profile D observations could arise from substitution, rejection, un-

	No event in analysis window	Disfluency or repair in analysis window
Higher suggestion correspondence	*Profile A*: response content corresponds with the suggestion, and no event occurs in the prespecified window. Consistent with smooth uptake, uncritical reuse, or missed event detection.	*Profile B*: response content corresponds with the suggestion and an event occurs in the prespecified window. Consistent with effortful uptake or concurrent production difficulty.
Lower suggestion correspondence	*Profile C*: response content has little correspondence with the suggestion and no event occurs. Consistent with independent production, missed support, or undetected correspondence.	*Profile D*: response content has little correspondence with the suggestion and an event occurs in the prespecified window. Consistent with substitution, rejection, unrelated planning difficulty, or floor management.

Table 1: Four behavioral profiles defined by suggestion correspondence and speech events in a prespecified analysis window. The cells summarize observable configurations and retain multiple candidate interpretations. Analyses should preserve both axes as continuous measures, and the quadrants are descriptive and support hypothesis generation.

Illustrated relation	Reconstructed illustration	Timing of speech events	Suggestion correspondence
Phrase match	"*um* like rolling suitcase"	Brief filled pause at retrieval onset and short response latency. Their relation to the AI event depends on the prespecified window and coding rule.	Surface and content correspondence with the reconstructed phrase suggested by AI.
Different content	"*um* I think we can include some basic items"	Filled pause at clause onset outside the prespecified analysis window.	Response contains content that differs from the reconstructed suggestion.
Match with repair	"*um* it's a pack-. . . packaged. . . packable backpack"	Filled pause before wording related to the suggestion, followed by truncation and self-repair.	Content correspondence with the AI's term ("packable") and phonological variation in its production.
Frame with substitution	"it's. . . *um* but maybe instead of a wool hat use a beanie"	Silent pause and filled pause before a lexical revision within the prespecified analysis window.	Syntactic frame correspondence with lexical substitution (wool hat → beanie).

Table 2: Four reconstructed illustrations of the component measures in Table 1. Content words were paraphrased into a different topical domain, and the original disfluency forms and positions were preserved. The rows illustrate measurement without validated profile assignments. Two trained annotators independently coded events against intervention timestamps and resolved disagreements through discussion. The source study received institutional review board approval (protocol 0149408). Italicized tokens mark filled pauses, and "…" marks silent pauses.

related planning difficulty, or floor management. A system comparison should use profile frequencies to select follow-up measures and should retain output quality, task performance, and participant reports as complementary outcomes.

Operationalizing the axes. Retain correspondence and speech timing as continuous measures in primary analyses, with the four profiles summarizing regions of their joint distribution. Any categorical cutpoint should be justified for the task, specified before analysis, and examined in a sensitivity analysis. This preserves information that universal high and low thresholds would discard.

Chance baseline for suggestion correspondence. Similar wording may have been likely without exposure to a particular suggestion. Compare each response with unshown suggestions matched on generator, task, topic, communicative function, length, and presentation format when possible. The displayed score, matched distribution, and their difference estimate how much correspondence is specific to the displayed content. Report unmatched dimensions because they limit the comparison.

Illustrative measurement check. The corpus first tests event detection and describes where events occurred. The source study recruited 38 speakers using English as a second language, and 35 recordings were usable for disfluency analysis. Two independently coded event intervals counted as a match when their temporal intersection divided by their union reached .25, yielding symmetric event-detection $F_1 = .981$. The adjudicated record contained 1,276 events with 2,286 category labels, of which filled pauses were the most fre-

quent category (1,039 labels). At least one event occurred in 62.3% of suggestion response intervals and 44.8% of clarification response intervals. The design included no unassisted condition, and the contexts differed in support function, response content, and observation time. This contrast therefore provides only a descriptive estimate of coverage.

The next check asks whether aggregate disfluency rate is associated with suggestion incorporation in this sample. Incorporation codes were available for 312 responses from two sources: 268 had graded codes and 44 had binary codes. Each source supplied a final code without independent duplicate coding, and the 22 positive binary cases lacked partial or full extent. Within participants, the probability that an incorporated response had a higher disfluency rate was .541 (exact $p = .347$, Holm-adjusted $p = .694$, 31 participants, 277 responses). This inconclusive association does not support using aggregate disfluency rate as a substitute for a separate correspondence measure.

The correspondence checks then ask whether the content measure behaves as intended. Using final corrected response spans, similarity increased with incorporation among 261 responses with graded codes and semantic scores (descriptive Spearman's $\rho = .483$). Across 332 responses, displayed suggestions were more similar than the mean of other suggestions of the same format that were not shown on that trial in 299 cases (90.1%), with mean similarity of .617 for displayed suggestions and .352 for comparisons (difference = .265). Both summaries contain repeated responses from participants and items. The alternative suggestions also concerned different items, so the second summary is a weak baseline that retains topic differences. These results establish event-detection feasibility and a descriptive relation between semantic correspondence and coded incorporation. The corpus has no measure of cognitive load aligned to individual events or independent labels for the profiles, leaving their cognitive interpretations for prospective tests.

3.4 Anticipated Objections

The most direct objection is that WER and surveys already cover system quality and user experience. They answer questions at different temporal scales: WER evaluates transcript accuracy, and surveys summarize retrospective experience. Analysis aligned to individual events asks where production changed around a particular intervention and whether its content entered the response. The framework is useful when this local process question matters to the system comparison.

A stronger objection concerns the nonspecificity of disfluency. Planning, retrieval, language background, and floor management can produce similar events. Temporal alignment and suggestion correspondence reduce this ambiguity, especially with comparisons within speakers or across matched episodes. Causal and cognitive claims still require the contextual and converging evidence specified in Table 3. This constraint defines the framework's evidential standard and prevents a behavioral profile from becoming a diagnostic label.

4 Application Across AI-MC Contexts

The framework earns practical value only when it changes how systems are compared. The following cases expose different ways that equivalent output scores can conceal process effects, and each case yields a distinct contrast for the two axes.

4.1 Proactive Content Scaffolding

Contextual suggestions and factual completions can improve the final response while imposing different demands during its production (Fang et al., 2025; Xia et al., 2023). Research on text input already documents changes in language use under reply and composition support (Arnold et al., 2020; Mieczkowski et al., 2021). In speech, relevant content available before response planning may reduce pauses at clause boundaries, whereas content that conflicts with an emerging plan may increase fillers, false starts, or repairs. Tests can separate these accounts by varying timing and semantic fit, comparing matched opportunities, and modeling individual baselines. Session trajectories then show whether an initial effect diminishes through adaptation or accumulates across interventions.

4.2 Lexical and Stylistic Adjustment

Word replacements, register adjustments, and tonal reframings intervene closer to formulation by supplying candidate lemmas and syntactic frames. Interactive alignment theory predicts lexical convergence through priming and shared expectations (Pickering and Garrod, 2004, 2013). Computational measures can compare convergence with the human partner and with the AI across turns (Nenkova et al., 2008; Reitter, 2014; Niederhoffer and Pennebaker, 2002). This comparison becomes

Requirement	Reference implementation and reportable choices
Unit and contrast	Analyze each suggestion opportunity and the participant's first substantive response. Report any unmediated condition, matched control trials, and adjustment for repeated observations within speakers.
Event anchors and windows	Record suggestion appearance, question completion, response onset, and response end. The illustrative corpus used four intervals: 10 seconds before appearance, appearance to question completion, question completion to response onset, and response onset to the earliest competing event, substantive partner turn, page transition, or 30 seconds after question completion.
Disfluency assignment	Assign an event to a window by onset. Report categories, counts, durations, denominator, annotation or extraction method, reliability, and timestamp sensitivity.
Content correspondence	Analyze only the participant's response text. Report surface overlap, semantic similarity, and incorporation codes separately. The illustrative check used `paraphrase-multilingual-mpnet-base-v2`. Report the model and segmentation rule.
Chance baseline	Compare the response with unshown suggestions matched on task, topic, communicative function, length, and presentation format. The illustrative corpus matched task and presentation format only. Report unmatched dimensions, the displayed score, comparison mean, and their difference.
Thresholds and missingness	Preserve continuous scores in primary analyses. Mark categorical profiles as exploratory. Record silence, abandoned turns, unavailable speech, and extraction failures separately.
Interpretation	Report language, task, turn position, floor management, and plausible alternative explanations. Use self-report, performance, multimodal measures, or experimental manipulation when making cognitive claims.

Table 3: Minimum reporting protocol. Values can vary across tasks and languages, and the listed information is required for comparison and interpretation.

consequential when accommodation is distributed unequally. Sociolinguistic research predicts greater accommodation from speakers with less positional power (Giles and Ogay, 2007). Higher suggestion correspondence and lower lexical diversity in that group would document an asymmetry requiring explanation. Establishing an AI effect then requires controls for power, topic, item content, and baseline language use.

4.3 Self-Presentation and Emotional Attunement

Tone refinement, empathic phrasing, and politeness support extend the same problem to self-presentation (Goffman, 1959; Walther, 1996). Suggestion correspondence localizes immediate reuse, and MATTR shows whether lexical diversity changes across the session. Comparisons within speakers across mediated and unmediated conditions can therefore test changes in expressive range. Participant and listener judgments remain necessary for assessing genuineness and appropriateness, outcomes that linguistic correspondence alone cannot determine. Across all three contexts, valid inference depends on following the relevant people, turns, and modalities beyond an isolated response.

5 Methodological Extensions

Extending the unit of analysis beyond one speaker and one turn introduces dependencies that the basic two-axis profile must represent explicitly.

5.1 Dyadic and Multiparty Settings

Dyadic AI-MC permits comparison of each speaker's speech rate, disfluency, and lexical choices across turns (Pickering and Garrod, 2004). Additional participants introduce overlap, floor holding, shifting addressees, and several potential sources of lexical convergence (Clark, 1996). Analyses should therefore model addressee, turn position, overlap, and group role. With these dependencies represented, the framework can test whether access to suggestions, content correspondence, and speech events are distributed asymmetrically across participants.

5.2 Speaker and Listener Roles

The analysis of speakers raises a corresponding question about what mediation changes for listeners. Automatic speech recognition captions, speech-to-speech translation, and speech language models may remove pauses and repairs from the signal (Wang et al., 2020; Rocholl et al., 2021; Défossez et al., 2024; Zhang et al., 2023; Skantze, 2021). Psycholinguistic studies show that filled pauses can direct attention toward novel or complex content and facilitate lexical integration (Arnold et al., 2003; Fox Tree, 2001; Corley et al., 2007), and listeners also adapt to a partner's disfluency patterns (Brennan and Schober, 2001). Smoothing may therefore improve surface fluency while reducing information used during comprehension.

This literature supports an untested prediction

about AI smoothing: removing natural disfluencies may reduce anticipatory cues and impair comprehension of novel material. A controlled study should manipulate filled pauses, silent pauses, and repairs separately. For fillers and silent pauses, lexical content and voice can remain constant while overall duration is balanced or modeled. Repair studies can compare matched fluent and repaired recordings and model duration differences. Recall, comprehension accuracy, clarification requests, echoed repetitions, and planning time in the listener's next turn capture consequences at different stages. This design could test whether an apparent improvement in output fluency creates a cost for the listener.

5.3 Multimodal Triangulation

The remaining ambiguity concerns what a localized speech event means. Complementary streams can connect the event to consultation, arousal, or broader acoustic change while preserving the limits of each measure.

Gaze and interaction logs. Gaze toward an AI display and interface logs can show that prompt consultation coincided with a speech event. Their inferential contribution is orientation to a region at the precision supported by the recording and coding method.

Physiological measures. Electrodermal activity (EDA) and heart rate variability (HRV) can provide converging measures of autonomic arousal (Taptiklis et al., 2023). Arousal remains compatible with deliberation, overload, and other explanations.

Acoustic features beyond disfluency. Speech rate, pitch range, and pause duration complement categorical counts. Extraction should use raw audio or a pipeline validated for retaining the target events.

Each stream addresses a different ambiguity, so the useful combination follows the claim being tested. Inferences about cognitive benefit or harm ultimately require performance, experience, or experimental outcomes aligned with the same episode.

6 Design Implications and Conclusion

Two systems can achieve the same WER, task completion, or satisfaction score while creating different production demands and different input for listeners. Disfluency measures locate observable production difficulty in time, and suggestion correspondence establishes how the response relates to system content. Their joint analysis makes these otherwise hidden differences available for system comparison.

The immediate use case is offline evaluation. Researchers can derive the measures through human coding, forced alignment, or ASR configured to retain disfluencies (Shriberg, 1994; Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021), with explicit validation of event retention, timestamps, and extraction error. Adaptive deployment sets a higher evidential threshold: a measure must prospectively predict benefit or difficulty in the target population, and responding to it must improve an outcome. An isolated disfluency spike or decline in lexical diversity cannot justify intervention.

The framework contributes four behavioral profiles, a minimum reporting protocol, and experiments that can test their interpretations for speakers and listeners. Its broader claim is that real-time AI-MC should be evaluated through the interaction that produces an output as well as through the output itself. This expanded unit of evaluation reveals when apparent system success coincides with observable production difficulty, response reuse, or fewer comprehension cues.

Limitations

The framework's value depends on disciplined interpretation of its boundaries. The illustrative corpus establishes event-detection and correspondence feasibility. Causal validation and independent labels for the four behavioral profiles require prospective data.

Pipeline dependence. Extracting fluency metrics requires human coding or a pipeline for ASR and disfluency detection. Outputs from many E2E speech models (Whisper, SeamlessM4T) and commercial ASR application programming interfaces (APIs) omit fillers and reparanda. Applying the framework therefore requires systems that preserve disfluencies (Wang et al., 2020; Jamshid Lou and Johnson, 2020; Rocholl et al., 2021), forced alignment over raw audio, or human annotation. Each approach introduces measurement error and cost that implementation work must quantify.

Speaker variability. Disfluency rates vary 5–10× across individuals as a function of age, language background, working memory capacity, and

clinical conditions (Bortfeld et al., 2001; Engelhardt et al., 2010). Contrasts within each speaker are preferred. Designs without an unmediated condition should use matched episodes, model speaker effects, and avoid unadjusted system rankings across speakers.

Limits of profile identification. The two axes are noisy observable measures. Similar response content can arise without suggestion use, and an event in the analysis window can reflect planning, retrieval, turn management, or measurement error. Chance baselines address prior content similarity, and contextual and multimodal evidence narrows the interpretations of speech events. The profiles remain descriptive and can be used to generate hypotheses.

Theoretical anchoring. Levelt's architecture organizes candidate production processes. Predictive processing and interactional accounts (Pickering and Garrod, 2013; Hagoort, 2019) motivate additional explanations. The proposed relations between disfluency types and production processes are probabilistic and require tests that account for context.

Operationalization. Table 3 specifies minimum reporting requirements. Lemmatizer choice, embedding model, n-gram window, MATTR window, and event windows still depend on language, register, and task. Comparisons across studies require reporting these choices and testing sensitivity to plausible alternatives.

Scope across languages. Our taxonomy and predictions are derived primarily from psycholinguistic literature on English. Filled-pause inventories, repair conventions, and cues at clause boundaries vary across languages, so the framework requires replication in each language before cross-language application.

Adaptive feedback. A system that responds to disfluencies may change how users speak over time. Periodic recalibration, control conditions that withhold intervention, and longitudinal evaluation are required to detect such effects.

Ethical Considerations

Speech behaviors have probabilistic associations with cognitive processes, creating risks in both data collection and interpretation.

Bias across speaker populations. Disfluency baselines vary across populations. Stuttering, aphasia, and Parkinson's-related dysarthria can produce disfluency unrelated to AI-MC, and speakers using a second language (L2) can differ from speakers using their first language (L1) (Segalowitz, 2010). Unadjusted application may conflate speaker characteristics with system quality. Designs that compare observations within each speaker, stratified analysis, and explicit baseline adjustment are prerequisites for population comparisons.

Surveillance dual use. Real-time extraction of speech features can support workplace monitoring. These measures should be limited to consented research evaluation. Any field deployment would require informed consent, bystander protection, data minimization protocols, and an explicit prohibition on inference about individual performance.

Privacy and institutional review. Applying the framework requires recording and processing speech, which is highly identifying personal data. Extracted features may retain sensitive information, and the source audio contains identity and content. Institutional review, secure storage, data minimization, and feature extraction on devices where feasible are appropriate safeguards. The recording used for Table 2 was collected under institutional review board protocol 0149408. The illustrations reconstruct the original events with content words paraphrased into a different topical domain.

Power asymmetry and equity. §4.2 motivates testing whether speakers with less positional power correspond more closely with AI suggestions. Such a pattern would raise a deployment concern because AI-MC may affect speakers unequally. Field evaluations should report outcomes by positional power and include pooled estimates only with the corresponding stratified results.

Feedback through training data. Training on speech processed through ASR, speech language models, or translation pipelines may change the distribution of disfluencies represented in future systems. Datasheets and provenance metadata should distinguish naturally produced speech from AI-MC speech.

Use of AI Assistants. We used large language model (LLM) writing assistants, including chat assistants and integrated development environment tools, to refine prose, copyedit text, and format LaTeX. The authors developed the research ideas, framework, analyses, citations, arguments, and bibliography and reviewed all assisted edits.

Materials Availability. Supporting materials can be found on the Open Science Framework at `https://osf.io/yc6dh/`.

References

Jennifer E Arnold, Maria Fagnano, and Michael K Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research*, 32(1):25–36.

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, pages 128–138. ACM.

Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 603–608, Tokyo, Japan. Association for Computing Machinery.

Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

Susan E Brennan and Michael F Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of memory and language*, 44(2):274–296.

Brian Butterworth. 1980. Evidences from pauses in speech, language production: Speech and talk, vol. i.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Martin Corley, Lucy J MacGregor, and David I Donaldson. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, 105(3):658–668.

Martin Corley and Oliver W Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R. Fussell. 2021. Bridging fluency disparity between native and nonnative speakers in multilingual multiparty collaboration using a clarification agent. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–31.

Paul E Engelhardt, Martin Corley, Joel T Nigg, and Fernanda Ferreira. 2010. The role of inhibition in the production of disfluencies. *Memory & Cognition*, 38(5):617–628.

Paul E Engelhardt, Joel T Nigg, and Fernanda Ferreira. 2013. Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta psychologica*, 144(2):424–432.

Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A wearable proactive ai" inner-voice" for contextual nudging. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, pages 1–9.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738.

Jean E Fox Tree. 2001. Listeners' uses of um and uh in speech comprehension. *Memory & cognition*, 29(2):320–326.

Howard Giles and Tania Ogay. 2007. Communication accommodation theory. In Bryan B. Whaley and Wendy Samter, editors, *Explaining Communication: Contemporary Theories and Exemplars*, pages 293–310. Lawrence Erlbaum, Mahwah, NJ.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Anchor Books, New York.

Frieda Goldman-Eisler. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, London.

Peter Hagoort. 2019. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58.

Robert J Hartsuiker. 2006. Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes*, 21(7-8):856–891.

Robert J Hartsuiker and Herman HJ Kolk. 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, 42(2):113–157.

Wanqing Psyche He and Susan R. Fussell. 2026. XPLAIN: A proactive scaffold across speech processing stages: Supporting non-native speakers in real-time ai-mediated turn-taking. In *Proceedings of the 8th ACM Conference on Conversational User Interfaces*, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific reports*, 13(1):5487.

Peter Indefrey and Willem JM Levelt. 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11).

Paria Jamshid Lou and Mark Johnson. 2020. Disfluency detection using a noisy channel model and a neural reranker. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5745–5755. ACL.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing*, 3rd edition. Draft chapters available at https://web.stanford.edu/~jurafsky/slp3/.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Willem JM Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38.

Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–14.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (ACL-HLT)*, pages 169–172. ACL.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Nazbanou Nozari, Gary S. Dell, and Myrna F. Schwartz. 2011. Is comprehension necessary for error detection? a conflict-based account of monitoring in speech production. *Cognitive psychology*, 63(1):1–33.

Claudy CE Oomen and Albert Postma. 2001. Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, 30(2):163–184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 61–74. MIT Press.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. ACL.

Albert Postma. 2000. Detection of errors during speech production: A review of speech monitoring models. *Cognition*, 77(2):97–132.

Peinuan Qin, Zicheng Zhu, Naomi Yamashita, Yitian Yang, Keita Suga, and Yi-Chieh Lee. 2025. AI-based speaking assistant: Supporting non-native speakers' speaking in real-time multilingual communication. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. ACL.

David Reitter. 2014. Alignment in web-based dialogue: Studies in big data computational psycholinguistics. *Topics in Cognitive Science*. Earlier formulation: Reitter, Keller & Moore (2011), "A computational cognitive model of syntactic priming," *Cognitive Science* 35.

Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency detection with unconstrained in-span operations. In *Proceedings of Interspeech 2021*.

Ardi Roelofs. 2020. Self-monitoring in speaking: In defense of a comprehension-based account. *Journal of Cognition*, 3(1):18.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Norman Segalowitz. 2010. *Cognitive bases of second language fluency*. Routledge.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Phd dissertation, University of California, Berkeley.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Nick Taptiklis, Merina Su, Jennifer H Barnett, Caroline Skirrow, Jasmin Kroll, and Francesca Cormack. 2023. Prediction of mental effort derived from an automated vocal biomarker using machine learning in a large-scale remote sample. *Frontiers in Artificial Intelligence*, 6:1171652.

Joseph B Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1):3–43.

Shaolei Wang, Wanxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Disfluency detection with unlabeled data and small annotated data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2162–2172. ACL.

Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. Crosstalk: Intelligent substrates for language-oriented interaction in video-based communication and collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773. ACL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Cadence-Stratified Multimodal Perception for Real-Time Child–Robot Interaction

Marwan Mohammed[1,2] **and Javad Zarrin**[2] **and Salma Elsayed**[2] **and Vera Kempe**[1]
[1]Faculty of Social and Applied Sciences, Abertay University, Dundee, UK
[2]Faculty of Design, Informatics and Business, Abertay University, Dundee, UK
m.mohammed@abertay.ac.uk

Abstract

Child–robot interaction places real demands on perception systems: a robot engaged in play with a child cannot afford to miss the child's focus of attention, yet running every perception module at every frame is computationally expensive on the kind of edge hardware that privacy-respecting robots must use. This paper addresses this tension by proposing a cadence-stratified multimodal perception pipeline, WordMate-CRI, in which each module runs at a rate that trades computation against staleness, informed by how long each cue remains stable for a potential interaction partner. Detection and tracking run continuously; action recognition uses rolling temporal buffers; emotion is refreshed periodically; and attributes such as age and gender are estimated once at the start of a session and cached. We evaluated WordMate-CRI through an LLM response-quality ablation study across six toy-play scenarios and a latency benchmark on Jetson AGX Orin hardware. The ablation shows that VLM enrichment and object evidence are the dominant grounding signals, with WordMate-CRI achieving a mean scene-grounded score of 0.967. On live Jetson input, cadence-based scheduling raises throughput from 4.3 to 18.2 frames per second (FPS), clearing the 15 FPS deployment target, while reducing WordMate-CRI latency from 6411.5 to 5844.5 ms. Taken together, the results make a practical case that deciding when each perception module should run matters just as much as deciding what it should recognise. This optimisation is important for systems aiming to produce contingent and timely child-directed speech in multi-modal contexts.

1 Introduction

Child–Robot Interaction (CRI) has become an important area in social, educational, therapeutic, and assistive robotics. In play, learning, communication support, healthcare, and autism-related interventions, robots are increasingly expected to act as responsive social partners for children (Scassellati et al., 2012; Kozima et al., 2007, 2009; Nadeem et al., 2025; Belpaeme et al., 2012; Efthymiou et al., 2022). Our project aims to design a robot capable of producing appropriate child-directed speech. This requires processing the attentional focus of the child in a timely and socially meaningful way, using multi-modal cues such as object detection, tracking of the child's position and movement, action, affect, engagement, and stable person-level attributes (Duncan et al., 2024; Cooper et al., 2026; Faria et al., 2025; Efthymiou et al., 2022).

A central challenge in real-time CRI is not only determining what information should be perceived, but when each perceptual process should run. Existing multimodal CRI and HRI systems combine vision, affective computing, behavioural analysis, and contextual reasoning (Duncan et al., 2024; Zhao et al., 2025; Belpaeme et al., 2012; Efthymiou et al., 2022). However, the temporal execution of these modules is often treated as an implementation detail rather than as an explicit design problem. Here we propose a more principled solution, recognising that multi-modal cues do not change at the same rate: object handling, position, and movement may change at faster rates than affect expression, which may evolve over a longer interval, while attributes such as age and gender remain more stable even in multi-child play sessions.

This temporal mismatch creates a practical cost. If every module is executed at every frame, the computation and latency increase, especially when several deep-learning models are used on the same robot. Recent HRI benchmarking also shows that current vision-language perception models still face difficult accuracy–latency trade-offs for real-time deployment (Shi et al., 2025). At the same time, reducing the update rate without a clear policy may cause the robot to miss relevant changes in the interaction. This paper therefore treats temporal scheduling as part of the perception architecture

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 22–32
October 29, 2026 ©2026 Association for Computational Linguistics

itself, rather than as a later optimisation step.

Our motivation is also consistent with evidence from psychology and neuroscience. Human perception is organised across multiple temporal scales: early sensory areas respond to rapidly changing input, while higher-level regions integrate information over longer windows (Hasson et al., 2008). A similar temporal structure has been shown during the comprehension of naturalistic stories (Lerner et al., 2011), event segmentation (Kurby and Zacks, 2008), adaptive temporal integration (Akyürek, 2025), and intrinsic neural timescales (Manea et al., 2022). These findings suggest a useful design principle for CRI: perceptual updates should be matched to the temporal stability of the to-be-processed cue.

Here we propose WordMate-CRI as a cadence-stratified multimodal perception architecture for real-time CRI. WordMate-CRI combines continuous detection and tracking, rolling-window action inference, periodic facial analysis, and cached demographic estimates. Section 3 specifies the update rules and their rationale.

Related work provides important foundations but does not directly address this specific scheduling problem. Multimodal HRI research highlights the need to coordinate heterogeneous perceptual channels in social environments (Duncan et al., 2024; Zhao et al., 2025). FlowAct shows the value of continuous perception and modular action subsystems (Dhaussy et al., 2024), while relevance-driven scheduling reduces latency by activating perception modules according to scene context (Huang et al., 2026). However, these approaches mainly schedule modules according to the relevance of the current scene. By contrast, the WordMate-CRI architecture assigns update cadences proactively, based on the expected temporal stability of the cues. Multiple-timescale and cognitive HRI models also show that robot processing need not be uniform in time (Yamashita and Tani, 2008; Umbrico et al., 2023, 2022), but their focus is mainly on reasoning, control, or cognitive organisation rather than module-level perception scheduling. The gap addressed in this paper is therefore a cue-stability-based temporal scheduling policy for multimodal child perception, where each module is updated according to the expected dynamics of the cue it estimates. Our main contributions are:

1. A cadence-stratified perception architecture of WordMate-CRI for real-time CRI, where perception modules run at different update rates instead of being executed uniformly at every frame.
2. An implemented multimodal perception pipeline combining continuous child detection and tracking, rolling-window action recognition, periodic emotion analysis, and sparse estimation of stable attributes.
3. A theory-to-engineering link between evidence for multi-scale temporal processing and a practical scheduling design for CRI.

This work is guided by two research questions: (RQ1) How does cadence-stratified scheduling affect visual-processing throughput and latency relative to per-frame processing? (RQ2) How does the availability of each nonverbal cue, namely gaze and head orientation for attention, affect for response tone, and object and action evidence for grounding, affect the quality of the generated response? Object detection and the VLM enrichment step that grounds response content operate at the segment level, evaluated once every five seconds, while the cadenced perception modules operate at the frame level; the throughput and latency gains reported in Section 5.2, addressing RQ1, arise from reducing frame-level computation, while response quality depends most on the segment-level evidence examined in Section 5.1, addressing RQ2. These are two distinct parts of the pipeline rather than a single measured trade-off curve. Thus, we evaluate processing throughput and latency against a per-frame baseline and separately examine how contextual information affects the generated response.

2 Related Work

The design draws on multi-timescale models of perception, real-time multimodal HRI, and child-centred robot perception.

A recurring finding in cognitive neuroscience is that the cortical areas integrate information over different time windows, with early sensory regions tracking fast input and higher-level regions accumulating evidence over longer periods (Hasson et al., 2008). Related patterns have been observed in story comprehension (Lerner et al., 2011), event segmentation (Kurby and Zacks, 2008), adaptive temporal integration (Akyürek, 2025), and intrinsic neural timescales (Manea et al., 2022). These findings suggest that different perceptual cues are naturally handled at different temporal resolutions,

a principle that has begun to surface in robotics as well. Yamashita and Tani demonstrated that functional hierarchy can emerge from a multi-timescale neural network in a humanoid robot setting (Yamashita and Tani, 2008), and related HRI architectures separate fast reactive processing from slower deliberative reasoning (Umbrico et al., 2023, 2022). Layered systems such as LAAIR extend this further by combining reactive and deliberative components within a single autonomous interactive robot (Jiang et al., 2018). What these works share is an acknowledgement that processing need not be temporally uniform. What they do not address is how individual social-perception modules in CRI should be scheduled, which is precisely the gap this paper targets.

On the multimodal perception side, the challenge of coordinating heterogeneous perceptual channels under real-world constraints is well documented (Duncan et al., 2024; Zhao et al., 2025). Systems such as FlowAct have moved away from strictly sequential pipelines by introducing a continuous asynchronous loop between perception and action (Dhaussy et al., 2024), and scheduling work has shown that activating perception modules according to scene relevance can significantly reduce latency (Huang et al., 2026). The tension between precision and latency is particularly sharp in edge hardware: studies of open-vocabulary mobile robot perception (Park et al., 2025) and the HRIBench evaluation of vision-language models (Shi et al., 2025) both confirm that current systems still struggle to meet the performance and timing demands of real-time deployment simultaneously. These studies make a strong case for principled scheduling, but their focus stays on general HRI, human–robot collaboration, or mobile robotics. Child-centred perception, with its specific cue profile and interaction demands, has so far received little direct attention. The two directly comparable systems differ from the approach proposed here in when the update rate is decided. The asynchronous perception-action system cited above keeps environment state continuously current and coordinates modular action subsystems through a planner, without assigning perception modules a fixed update rate in advance. The relevance-driven scheduler cited above activates modules based on prior outputs and scene context, which is a reactive, relevance-triggered policy rather than a cadence set proactively from the expected stability of each cue. The schedule proposed in this paper is therefore distinguished by assigning per-cue update rates ahead of runtime, evaluated specifically in a child-perception setting rather than in general human-robot interaction. That gap is only partly filled by the CRI literature. Early work established the therapeutic value of social robots for children, including simple platforms such as Keepon in autism-related settings (Scassellati et al., 2012; Kozima et al., 2007, 2009), and recent reviews confirm growing interest in robots that support communication, learning, and social behaviour in children with autism spectrum disorder (Nadeem et al., 2025). More recently, multimodal CRI has increasingly been framed as a socially situated process that requires coordinated perception and timely response (Belpaeme et al., 2012). For example, ChildBot demonstrated multi-robot perception and interaction with children at scale (Efthymiou et al., 2022); affective computing and biophysical signals have been integrated to make interactions emotionally aware (Faria et al., 2025), and incremental visual learning has been explored to handle child-specific appearance variation (Efthymiou et al., 2021). Engagement detection ties many of these threads together, drawing on gaze, head orientation, action, and affect simultaneously (Lu et al., 2024; Kim et al., 2023). However, what this body of work focuses on is *what* should be recognized and *how* recognition should be performed. The question of *when* each perceptual module should run during live interaction has received far less attention. This is particularly important for systems that aim to produce child-directed speech contingent on children's focus of attention, thereby supporting word learning (Yu et al., 2019; Schroer and Yu, 2023). Rather than proposing new recognition models or a new multimodal CRI platform, we explore a cue-stability-based temporal scheduling policy: continuous updates for detection and tracking, rolling-window inference for behavioural cues, periodic refresh for affect, and sparse session-level estimation for stable attributes such as age and gender.

3 Proposed Pipeline

3.1 WordMate-CRI Overview

In the proposed cadence-stratified multimodal perception architecture each perception module operates at an update rate matched to the temporal stability of each cue. The system processes a continuous RGB-D video stream and produces temporally aggregated scene representations that drive

spoken language responses generated by a locally deployed LLM. The WordMate-CRI is structured across three computational tiers operating at distinct temporal granularities (Figure 1):

1. Tier A—Frame loop: person and object detection (YOLO) and multi-person tracking (Norfair) run on every processed frame. Action recognition (R(2+1)D-18) uses a rolling temporal buffer with stride-based inference. Head orientation (6DRepNet) is computed in the frame loop at its configured stride.

2. Tier B-Cadenced: asynchronous background workers whose update rate trades computation against cue staleness rather than assuming the underlying cue changes slowly. Facial affect and gaze (OpenFace3) refresh every 5 person-detection frames (every frame when head-pose is enabled), and child age and gender (MiVOLO) are polled every 8 person-detection frames and locked after 4 successful inference runs; both run as background workers whose results are collected within the same frame call.

3. Tier C-Segment-level (5-second windows): Temporal aggregation of Tier A and B observations into a structured scene record, parallel VLM enrichment on the segment keyframe and spoken response generation by the LLM.

Detection and tracking are updated continuously to follow object and person movement, while action inference integrates evidence over a temporal window. Facial analysis is refreshed periodically to reduce repeated computation, and demographic estimates are cached while the tracked person remains unchanged. Gaze-to-object association is updated each frame using the available estimates. Although the transitions in affect, gaze, and head pose are fast, the relevant cue values persist considerably longer: spontaneous facial expressions typically last 0.5 to 4 s, and looks to objects during toy play extend over seconds (Yu et al., 2019). At the deployed refresh interval of approximately 275 ms, each such event is sampled at least twice before it ends.

A natural question is whether the modular pipeline could be replaced by a single locally deployed VLM. This approach was deliberately avoided for four reasons: (1) privacy constraints preclude cloud-based VLMs since transmitting children's video violates GDPR and COPPA and diminishes user trust; (2) accuracy limitations of lightweight local VLMs (e.g., LLaVA-1.5-7B, Phi-3-Vision) make them unreliable for child toy-play content, where objects and actions are underrepresented in standard training corpora; (3) latency incompatibility, as 7B+ parameter models require 500 ms–2 s per inference on Jetson AGX Orin, yielding frame rates below our adopted 15 FPS visual-processing target (see Section 4.3.3); and (4) hallucination rates in edge-deployed VLMs are difficult to calibrate, unlike the modular classifiers which offer bounded, characterisable uncertainty. The VLM is therefore retained solely as a segment-level scene-enrichment oracle.

3.2 Segment Aggregation and LLM Response Generation

At the segment level (Tier C), outputs from the cadenced perception modules are aggregated over a 5-second window into a structured context record. The aggregation is signal-specific: actions are selected by majority vote from the R(2+1)D-18 rolling-buffer classifier, with an action stability score used to flag possible activity switching when the dominant label covers less than 50% of the window. Emotion is summarised using the dominant OpenFace3 label, an expression distribution, and an optional trajectory field when the dominant emotion changes between the first and second half of the segment. YOLO object detections are deduplicated by label, ranked by average confidence, and filtered using a 30% presence-ratio threshold to remove transient detections. Gaze behaviour is represented through a focus ratio and the number of unique gaze targets, allowing the system to distinguish sustained attention from scattered exploration. Age and gender estimates from MiVOLO are aggregated over four successful inference runs, scheduled every eight person-detection frames, and then cached until the demographic lock is reset.

The resulting context record is used to generate the robot's spoken response. A VLM (Gemma3-4B via Ollama) runs in parallel on the segment keyframe to provide high-level toy and activity hints, while a locally deployed LLM (Qwen2.5-3B-Instruct via Ollama) produces a four-line caregiver-style response: (1) a warm observation, (2) a gentle invitation, (3) a playful question, and (4) an encouraging follow-up. The final text is converted to speech using Piper TTS and delivered by the robot.

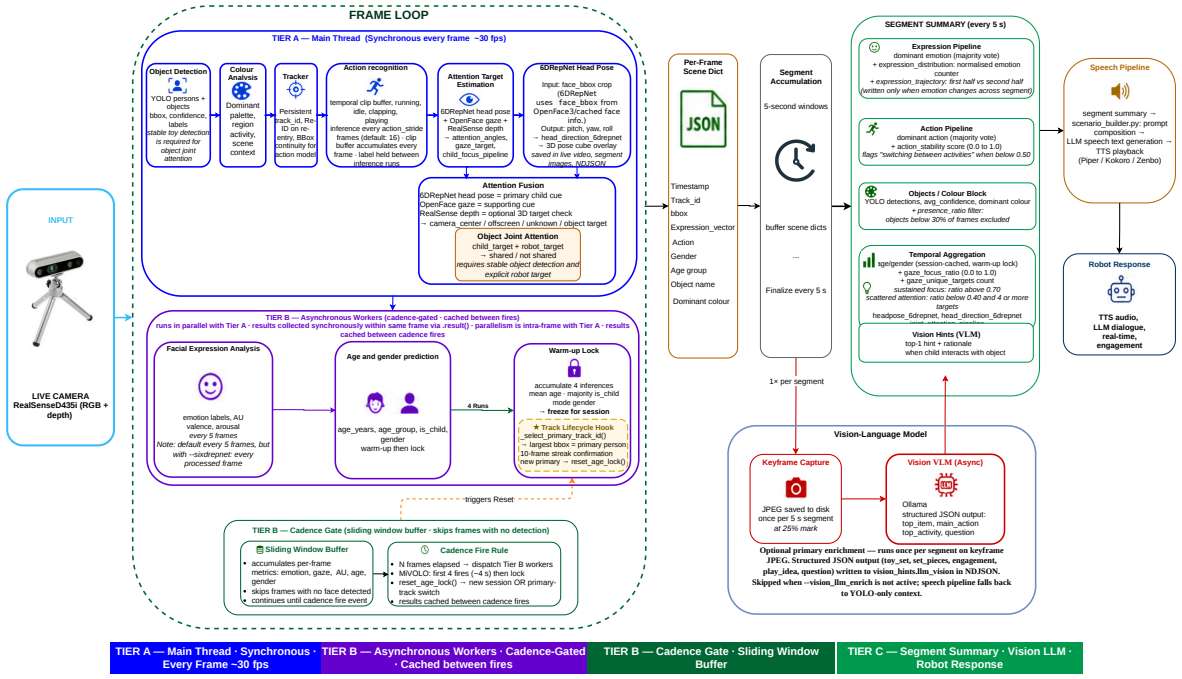


Figure 1: WordMate-CRI. Tier A (main thread, every frame) performs person detection, multi-person tracking, action recognition (rolling buffer with a 16-frame stride in the reported deployment benchmark), and head orientation estimation. Tier B (asynchronous background workers) refreshes facial affect and gaze every 5 person-detection frames and estimates age and gender every 8 person-detection frames, locking the latter after 4 successful inference runs. Tier C aggregates Tier A and B outputs once per 5-second segment, triggering VLM scene enrichment and spoken response generation.

4 Materials and Experimental Setup

4.1 Hardware

The pipeline ran on an NVIDIA Jetson AGX Orin 64GB Developer Kit as the main onboard compute unit. Visual input came from an Intel RealSense D435i RGB-D camera, providing aligned colour and depth streams for scene understanding and gaze-to-object association. Responses were played through a USB stereo speaker. Offline development and evaluation used a Windows workstation with an NVIDIA RTX 5000 Ada Generation Laptop GPU.

4.2 Software and AI Framework

We implemented the pipeline in an isolated Conda environment to support reproducible execution across both platforms. Visual perception combined YOLO object detection, Norfair tracking, R(2+1)D-18 action recognition, OpenFace3 facial analysis, 6DRepNet head-pose estimation, and a custom RGB-D gaze-to-object association module. MiVOLO provided demographic estimates after a short warm-up period. Responses were generated via Qwen2.5-3B-Instruct through Ollama, vision-language enrichment by Gemma3-4B, and spoken output synthesised using Piper TTS.

4.3 Evaluation Design

4.3.1 Dataset and Settings

The response evaluation uses six proxy toy-play clips depicting block building, cup stacking, picture-book interaction, toy-kitchen play, doctor-kit play, and racetrack play. Each clip is 10 seconds long at 30 FPS and 640 × 480 resolution, giving 60 seconds of footage. Two approximately five-second windows per clip yield 12 scene segments. Evaluating each segment under eight named conditions produces 96 response records. Child demographic context is assigned for response generation and is not a measurement of the performers' ages. Recorded benchmarking processes saved video, whereas live benchmarking processes camera input on the Jetson AGX Orin and Windows workstation. The live evaluation contains five segment records per Jetson condition and three baseline/four proposed records on Windows. These counts describe timing records, not independent participants or matched repetitions of the same scene.

4.3.2 Ablation Study

The response study examines how changes to the available perceptual context affect scene-grounded caregiver-style language. Each generated response is evaluated against the same human-reviewed ground-truth reference for its scene segment across all conditions. This ground-truth reference comprises a verified scene description, an example of a contextually appropriate child-directed response for that scene, and scene keywords.

Ground-Truth: Ground truth was created from the output generated by running the WordMate-CRI across all six recorded clips. For each toy segment, saved frames, segment NDJSON files, VLM hints, generated responses, and prompt/response tables were manually reviewed. The ground-truth JSON was then refined where needed so that the scene description, reference caregiver response, and keywords reflected what was visible in the scene. This was necessary because automatic detections sometimes produced noisy labels, such as identifying colourful blocks as a skateboard or a bottle, which were not present in the scene. The human-reviewed ground truth was therefore used as the verified scene reference. It included a caregiver-style reference response, written as a warm observation, gentle invitation, playful question, and encouraging follow-up, together with scene keywords such as the toy object, action, and play activity. The ground-truth reference for each segment was initialised automatically from the recorded scene output and then manually reviewed against the corresponding video, saved frames, and VLM hints. Fields that visibly conflicted with the clip were corrected. These human-reviewed references were then used to evaluate the system's generated responses.

Ablation Conditions: Each segment was evaluated under the WordMate-CRI and seven ablated conditions. WordMate-CRI retained all available segment information and served as the baseline. There are seven ablation conditions: no action recognition (-Action), no emotion detection (-Emotion), no gaze and focus-target information (-Gaze), no age and gender (-Age/Gender), no VLM scene hints (-Vision LLM), no object detections (-Object), and retention of solely object detections (Obj. only). These conditions were used to test how each signal affected robot response based on scene recognition.

Five of the seven conditions, namely no action recognition, no emotion detection, no gaze and focus-target information, no age and gender, and no object detections, remove one perceptual signal at a time in order to isolate its marginal contribution to the generated response, following a leave-one-out design. The remaining two conditions test the extremes of visual-semantic grounding, directly motivated by the design question raised in Section 3.1 of whether a modular pipeline is preferable to interpretation based solely on the vision-language model: the no-VLM-scene-hints condition removes that enrichment step entirely, while the objects-only condition retains solely object-detection evidence. Because VLM scene hints are themselves partly derived from detected objects, the no-object-detections condition removes both the object list and the associated VLM hints together, and is therefore best read as a combined removal of object-derived evidence rather than an isolated test of object detection alone.

Evaluation Metrics: Each condition was evaluated using four complementary metrics. The Scene Grounded Score (SGS) was used as the primary measure of response quality and scene faithfulness, while SGS degradation, hallucination rate, and Keyword Hit Rate (KHR) were used to examine grounding loss, error patterns, and lexical overlap. SGS is an automated score, normalised between zero and one, computed from five weighted components, namely object grounding, action grounding, scene consistency, developmental-language appropriateness, and caregiver-response format, each contributing up to two points, from which a hallucination penalty of up to two points and a generic-language penalty of up to one point are subtracted; it reflects the alignment between a generated response and the verified scene and response format, and is not a validated measure of child language development. KHR is a particularly important parameter that ensures correct labelling responses by the robot to promote children's word learning, as it is the proportion of reference scene keywords, matched through lemma-level overlap, that appear in the generated response, a lexical-coverage measure rather than a semantic one, reported alongside SGS rather than in its place. Correct object and action labelling in caregiver speech is the mechanism by which KHR connects to word learning (Yu et al., 2019; Schroer and Yu, 2023).

The ablation study in this section was run with the implementation's default action-recognition stride of 8 frames, distinct from the stride of 16

Metric	Description
Per-frame processing time	Defined as the mean duration of `analyze_frame()` over a segment, measured by $\mathrm{T1\_mean\_frame\_ms}$. It indicates the frame-level cost of visual processing.
Frames per second (FPS)	Calculated as $1000/\mathrm{T1\_mean\_frame\_ms}$. It measures visual-processing throughput and is used to assess performance against the 15 FPS Jetson live target.
Vision processing time	Defined as $\mathrm{vision\_proc\_ms} = T2 - T0$, measuring the wall-clock time from the first frame of the segment to segment finalisation.
Speech pipeline latency	Defined as $\mathrm{pipeline\_ms} = T7 - T_{\mathrm{seg\_read}}$, measuring the delay from segment pickup to the first spoken word.
System latency	Defined as $\mathrm{system\_latency\_ms} = \mathrm{vision\_proc\_ms} + \mathrm{pipeline\_ms}$. It summarises visual processing and speech-generation latency while excluding robot turn-taking time.

Table 1: Evaluation metrics used in Latency analysis.

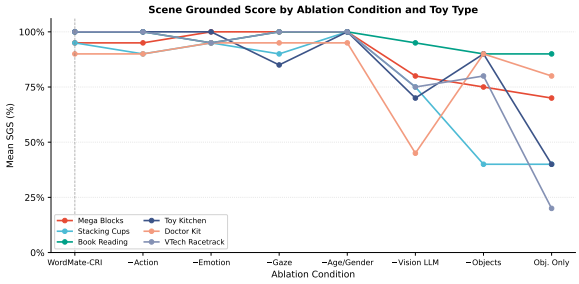


Figure 2: Scene Grounded Score across ablation conditions for each toy type.

frames used in the latency benchmark of Section 5.2; both reflect the same rolling-buffer mechanism at different sampling rates.

4.3.3 Latency Benchmark: Per-Frame Baseline versus WordMate-CRI

The timing study compares the per-frame baseline with WordMate-CRI using the recorded and live input settings described above. Table 1 defines the five reported timing measures. We deployed 15 FPS as a practical target for live visual-processing throughput on the Jetson, informed by (Park et al., 2025), who identified rates above 15 FPS as suitable for balancing responsiveness and perceptual capability in general HRI. This is an engineering benchmark for our deployment, rather than a validated minimum for child–robot interaction.

5 Results and Discussion

The ablation study showed that VLM scene enrichment and object evidence contributed most strongly to scene-grounded response quality (Section 5.1). Alongside these response-quality findings, the latency benchmark showed that cadence-based scheduling improved visual-processing throughput and reduced system latency, enabling live Jetson deployment to exceed the adopted 15 FPS target (Section 5.2). However, because response quality was not directly compared between per-frame and cadenced processing, these results do not establish accuracy preservation.

5.1 Ablation Study Results

The SGS results in Figure 2 show that WordMate-CRI (Full) consistently achieved strong scene-grounded response quality across all six toy scenarios, with a mean SGS of 0.967 across the twelve evaluated segments. Removing action, emotion, gaze, or demographic information produced only minor changes, with mean SGS values of 0.958, 0.967, 0.950, and 0.992 respectively, suggesting these modalities had limited effect on core scene recognition, though they may still shape response detail. Larger reductions appeared when visual-semantic grounding was weakened: removing VLM context dropped the mean SGS to 0.733, removing object detections to 0.775, and the objects-only condition produced the lowest score at 0.567. This confirms that high-level visual interpretation and object evidence are complementary — object labels contribute useful grounding, but are not sufficient on their own for a contextually appropriate robot response.

The hallucination-rate Figure 3 results help explain the SGS reductions observed in the more severe ablations. When scene grounding was weakened, the model did not simply become less specific; in several cases it shifted towards an entirely different but plausible play scenario. In the Mega Blocks clips, for instance, the verified activity was block stacking, yet some ablated responses referred to a skateboard or bottle, likely because the colourful block structure was mislabelled by lower-level detectors. Without VLM hints to correct this, the model followed the noisy cues. Similar drift was observed across other toys: stacking cups responses moved towards ball play, toy kitchen clips occasionally shifted to ball rolling, and racetrack clips drifted towards cups, bubbles, or cooking-related language. Although these responses remained plausible within a child-play setting, they were not faithful to the verified scene.

The KHR shown in Figure 4 supports the SGS findings but also reveals why lexical overlap alone is an unreliable evaluation measure. The full pipeline sometimes received a lower KHR than an ablated condition because KHR rewards only

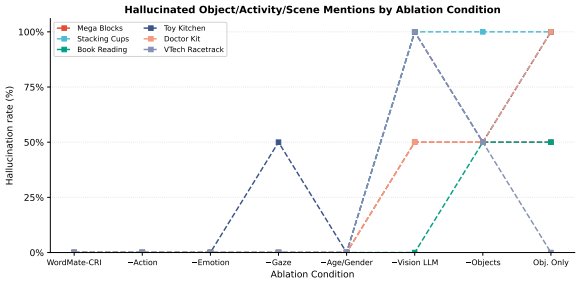


Figure 3: Rate of hallucinated object, activity, or scene mentions by ablation condition and toy type. Each toy-condition combination is computed from two evaluated responses, so plotted values fall at 0%, 50%, or 100%; they reflect rule-based hallucination flags rather than a fine-grained or exhaustively human-annotated rate.

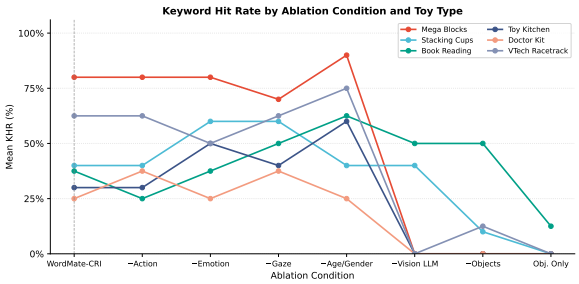


Figure 4: Keyword Hit Rate across ablation conditions for each toy type.

exact or normalised keyword matches. A contextually appropriate response using “pieces”, “goes on top”, or “make it taller” scores lower than one that mentions “blocks” by chance while describing the wrong activity entirely. KHR can also dip in the full condition when noisy automatic detections — an incorrect gaze target or a mislabelled object — are included in the prompt and the model partially follows them. Removing that noisy field can therefore produce a cleaner keyword match without improving actual scene recognition. Such occasional improvements under ablation should not be read as evidence that the removed modality is harmful; they point instead to the need for confidence weighting and conflict resolution between low-level detections and VLM interpretation.

Overall, the ablation results show that the pipeline benefits from multi-source visual context, though modalities contribute unequally. VLM enrichment provides the strongest grounding signal, and its removal causes the largest quality drop. Object detections add value, but mainly when interpreted alongside wider scene context. The uneven contribution of some modalities likely reflects the fact that the underlying perception models were not trained for young children, caregiver–child interaction, or toy-rich settings, which contributes to noisy detections and occasional grounding errors. These findings highlight the importance of confidence weighted prompt construction and better conflict resolution between low-level detections and VLM interpretation as clear priorities for future work.

5.2 Latency Results

Figure 5 summarises latency across both platforms, Windows and Jetson, using recorded-video and live-camera inputs, covering offline processing behaviour and real-time deployment performance.

In the recorded video, the WordMate-CRI improved mean FPS from 7.2 to 29.7 on the Jetson and reduced system latency from 21.90s to 8.13s. A similar trend was observed on Windows, where FPS increased from 7.0 to 17.6 and system latency decreased from 24.14s to 10.88s. These results indicate that the cadence strategy removes unnecessary per-frame computation across both platforms. The stage-level results show that the main savings come from frame processing and vision-side computation. Other stages, particularly LLM generation and TTS, still contribute noticeably to overall delay, especially when responses are longer or speech synthesis takes more time to produce the first line. Therefore, the WordMate-CRI should be understood as a visual scheduling improvement that makes the interaction pipeline more practical on edge hardware, rather than a complete solution to all latency sources.

The live-camera provides the most important deployment evidence. On the Jetson AGX Orin, the per-frame baseline achieved only 4.3 FPS, whereas the WordMate-CRI reached 18.2 FPS, exceeding the 15 FPS live target for robot deployment. The same improvement is reflected in per-frame processing time, which dropped from 235.9 ms to 60.0 ms. This shows that the WordMate-CRI strategy changes the live system from a low-frame-rate pipeline into one that can maintain adequate temporal sampling during child play. Windows live input also improved from 2.0 to 9.5 FPS, though it did not reach the 15 FPS target, suggesting the Windows live-camera path remains more variable.

The speech pipeline results are more nuanced. On Jetson live input, speech pipeline latency was 1133.9 ms under the WordMate-CRI, compared with 1021.7 ms for the baseline. This slight increase does not undermine the cadence strategy, as the method is designed to reduce vision-side

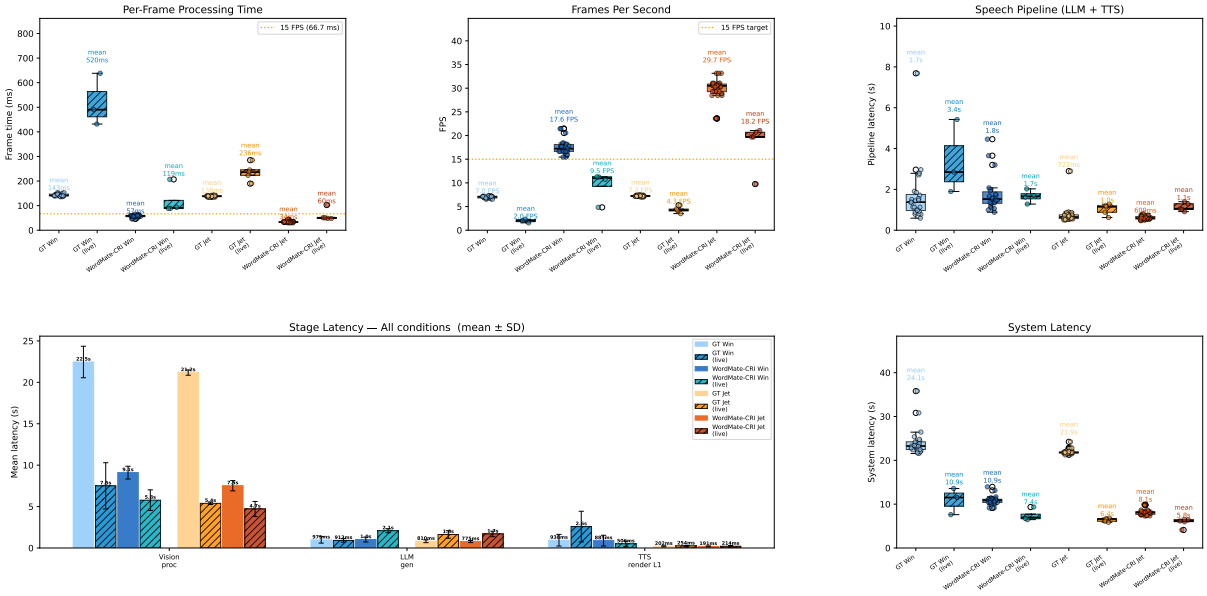


Figure 5: Latency for the per-frame baseline and WordMate-CRI on Windows and Jetson platforms. WordMate-CRI improves visual-processing throughput and enables the Jetson live deployment to exceed the 15 FPS target, while system latency is reduced despite remaining LLM and TTS costs. FPS is averaged from the reciprocal of each segment's mean frame time. Live-condition sample counts are five records per Jetson condition and three (baseline) or four (proposed) per Windows condition; stage-latency bars show mean plus or minus one standard deviation

computation rather than shorten LLM or TTS generation. Importantly, the WordMate-CRI keeps the post-segment response delay close to one second while greatly improving live frame throughput.

On Jetson live input, system latency decreased from 6411.5 to 5844.5 ms. The smaller reduction relative to the FPS gain reflects the fixed real-time segment window: the cadence strategy does not make a five-second segment arrive earlier, but it allows the system to process it with substantially less computational pressure. The Jetson live result is therefore the key deployment finding, with the recorded-video benchmark providing supporting evidence across the six toy clips.

6 Conclusion

Building a robot that keeps up with a child's attentional focus during play is as much a scheduling problem as it is a recognition problem. Not all information a robot needs to perceive changes at the same speed, and treating all cues as if they deserved equal computational attention is wasteful on the edge hardware that real deployments depend on. WordMate-CRI increased live Jetson throughput from 4.3 to 18.2 FPS. In the separate response evaluation, the full-context condition achieved a mean scene-grounded score of 0.967. That said, the perception models were not built with children or toy-play in mind, and LLM generation and speech synthesis still carry latency the cadence strategy does not address. Future work should bring the perception models closer to child-centred environments and develop principled conflict resolution between low-level detections and VLM interpretation. What the results establish clearly is that temporal scheduling deserves to sit alongside model selection and multimodal fusion as a first-class design decision in child-robot perception.

6.1 Limitations

This study is an initial feasibility evaluation based on a small set of controlled toy-play scenarios. The current pipeline does not yet process children's speech or other auditory input; augmenting the pipeline accordingly in the future will add further important information about children's communicative intentions. Most of the perception components in the pipeline are pretrained models, including YOLO for object detection, R(2 + 1)D-18 for action recognition, OpenFace3 for gaze and facial expression analysis, 6DRepNet for head pose and MiVOLO for age and gender estimation. These models were not specifically trained for young children, caregiver–child interaction, or toy-rich play settings, which may explain some noisy detections and grounding errors. Furthermore, reference re-

sponses were reviewed by a single annotator who also inspected the system's generated outputs during the same pass, so the evaluation is not blind to model output, and no independent inter-annotator check was performed. Future work should collect child–caregiver interaction data to refine the action model and adapt object detection to toys and materials commonly found in child-centred environments. Larger and more diverse data would also support confidence weighting, conflict resolution between low-level detections and VLM hints, and stronger evaluation in real-world CRI settings. Finally, although WordMate-CRI reduces vision-side computation, LLM generation and text-to-speech remain sources of latency that require further optimisation.

Acknowledgments

The authors wish to acknowledge generous funding by the Northwood Charitable Trust.

References

Elkan G. Akyürek. 2025. Temporal integration as an adaptive process in visual perception, attention, and working memory. *Neuroscience & Biobehavioral Reviews*, 170:106041.

Tony Belpaeme and 1 others. 2012. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53.

S. Cooper, B. Pou, A. Mayoral-Macau, A. Redondo, D. Ríos, and R. Ros. 2026. EMOROBCARE: A low-cost social robot for supporting children with autism in therapeutic settings. In *Social Robotics + AI, ICSR+AI 2025*, volume 16131 of *Lecture Notes in Computer Science*, pages 607–615. Springer, Singapore.

Timothée Dhaussy, Bassam Jabaian, and Fabrice Lefèvre. 2024. FlowAct: A proactive multimodal human-robot interaction system with continuous flow of perception and modular action sub-systems. *arXiv preprint arXiv:2408.15864*.

J. A. Duncan, Farshid Alambeigi, and Mitch W. Pryor. 2024. A survey of multimodal perception methods for human–robot interaction in social environments. *ACM Transactions on Human-Robot Interaction*, 13(4):1–50.

Nikolaos Efthymiou, Panagiotis P. Filntisis, Petros Koutras, Anna Tsiami, James Hadfield, Gerasimos Potamianos, and Petros Maragos. 2022. ChildBot: Multi-robot perception and interaction with children. *Robotics and Autonomous Systems*, 150:103975.

Nikolaos Efthymiou, Panagiotis P. Filntisis, Gerasimos Potamianos, and Petros Maragos. 2021. Visual robotic perception system with incremental learning for child–robot interaction scenarios. *Technologies*, 9(4):86.

Diego R. Faria, A. L. Godkin, and P. P. da Silva Ayrosa. 2025. Advancing emotionally aware child–robot interaction with biophysical data and insight-driven affective computing. *Sensors*, 25(4):1161.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.

D. Huang, X. Zhang, and K. Youcef-Toumi. 2026. Perceive what matters: Relevance-driven scheduling for multimodal streaming perception. *arXiv preprint arXiv:2603.13176*.

Yuqian Jiang, Nick Walker, Minkyu Kim, Nicolas Brissonneau, Daniel S. Brown, Justin W. Hart, Scott Niekum, Luis Sentis, and Peter Stone. 2018. LAAIR: A layered architecture for autonomous interactive robots. *arXiv preprint arXiv:1811.03563*.

Y. Kim, H. Chen, S. Algohwinem, C. Breazeal, and H. W. Park. 2023. Joint engagement classification using video augmentation techniques for multi-person HRI in the wild. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pages 698–707, London, United Kingdom.

Hideki Kozima, Marek P. Michalowski, and Cocoro Nakagawa. 2009. Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, 1:3–18.

Hideki Kozima, Cocoro Nakagawa, and Yuriko Yasuda. 2007. Children–robot interaction: A pilot study in autism therapy. In *Progress in Brain Research*, volume 164, pages 385–400.

Christopher A. Kurby and Jeffrey M. Zacks. 2008. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2):72–79.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

S.-R. Lu, J.-H. Lo, Y.-T. Hong, and H.-P. Huang. 2024. Implementation of engagement detection for human–robot interaction in complex environments. *Sensors*, 24(11):3311.

Ana M. G. Manea and 1 others. 2022. Intrinsic timescales as an organizational principle of neural processing across the whole rhesus macaque brain. *eLife*, 11:e75540.

Muhammad Nadeem, J. M. H. Barakat, D. Daas, and A. Potams. 2025. A review of socially assistive robotics in supporting children with autism spectrum disorder. *Multimodal Technologies and Interaction*, 9(9):98.

J. Park, P. Kim, and D. Ko. 2025. Real-time open-vocabulary perception for mobile robots on edge devices: A systematic analysis of the accuracy-latency trade-off. *Frontiers in Robotics and AI*, 12.

Brian Scassellati, Henny Admoni, and Maja Matarić. 2012. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294.

Sara E. Schroer and Chen Yu. 2023. Looking is not enough: Multimodal attention supports the real-time learning of new words. *Developmental Science*, 26(2):e13290.

Z. Shi, E. Zhao, N. Dennler, J. Wang, X. Xu, K. Shrestha, M. Fu, D. Seita, and M. Matarić. 2025. HRIBench: Benchmarking vision-language models for real-time human perception in human-robot interaction. In *Proceedings of the 19th International Symposium on Experimental Robotics*.

Alessandro Umbrico, Amedeo Cesta, Francesca Fracasso, Andrea Orlandini, and Gabriella Cortellessa. 2022. Shaping cognitive control for HRI through the dual process theory. In *Proceedings of the Workshop on Cognitive Architectures for HRI*, volume 3332 of *CEUR Workshop Proceedings*.

Alessandro Umbrico, Riccardo De Benedictis, Francesca Fracasso, Amedeo Cesta, Andrea Orlandini, and Gabriella Cortellessa. 2023. A mind-inspired architecture for adaptive HRI. *International Journal of Social Robotics*, 15:371–391.

Yuichi Yamashita and Jun Tani. 2008. Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Computational Biology*, 4(11):e1000220.

Chen Yu, Sumarga H. Suanda, and Linda B. Smith. 2019. Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science*, 22(1):e12735.

W. Zhao, K. Gangaraju, and F. Yuan. 2025. Multimodal perception-driven decision-making for human-robot interaction: A survey. *Frontiers in Robotics and AI*, 12:1604472.

Multimodal Code-Mixed Hindi-English Dataset For Humor Recognition in Conversations

Abhishek Singh Bisht[1], Rakesh Kumar[1], Amrita Chaturvedi[2]
[1]Gautam Buddha University, Greater Noida, India
[2]Indian Institute of Technology (BHU), Varanasi, India
bishtmlai@gmail.com, rakesh.kumar@gbu.ac.in, amrita.cse@iitbhu.ac.in

Abstract

Humor recognition in conversations is a highly challenging task due to its reliance on context, sarcasm, timing, and multimodal cues. While several humour datasets exist for English, resources for Indic languages like Hindi remain little scarce, limiting the development of culturally aware conversational AI systems. To address this gap, we introduce a new multimodal dataset for humour detection from conversational dialogues. The dataset consists of 5203 annotated utterances of 2 season containing 8 episodes each from a popular Indian web series *"Panchayat"*. Utterances are kept in their original Devanagari script as spoken in the series, without transliteration or standardization into pure Hindi, to retain the natural conversational tone and cultural nuances. Each utterance is annotated with a binary humour label (Humour / Not Humour) and includes aligned textual, acoustic and visual information, making it suitable for multimodal and context-aware humour research. Our experiments highlight the complementary nature of multimodal signals and the importance of conversational context in understanding Hindi humour. This dataset is expected to promote research on computational humour, sarcasm detection and multimodal dialogue understanding for low-resource languages. The dataset and baselines are available at https://huggingface.co/datasets/Abhis4e/Panchayat-multimodal-dataset and https://github.com/singhbishtabhishek/Multimodal-Dataset.

1 Introduction

Humor is a subtle and context dependent aspect of human communication that is often difficult to detect automatically (Kalloniatis and Adamidis, 2025; Purandare and Litman, 2006). In conversational settings humor is not conveyed through text alone but is also influenced by tone of speech, timing and speaker interactions (Bertero and Fung, 2016; Lee et al., 2016). Relying only on textual content can miss important cues especially in dialogues where the same sentence may or may not be humorous depending on delivery of speaker. This makes humor detection particularly challenging in real world scenarios where multiple modalities like text, audio and visual signals jointly contribute to signal a dialogue as humourous or not (Hasan et al., 2019; Tsai et al., 2019; Baltrušaitis et al., 2019).

This challenge is amplified in Hindi conversational data which is typically informal, noisy(spoken) and reflective of natural spoken usage rather than standardized language. Utterances often include mixed vocabulary, transliterated expressions and speaker-specific styles, making the task more complex than working with clean text (Chauhan et al., 2021, 2022). These challenges motivated us to investigate humor detection in real-world Hindi conversational data using a multimodal dataset.

For this study we introduce a multimodal Hindi conversational dataset consisting of code-mixed utterances written in Devanagari script along with textual dialogues, audio and visual modalities. The dataset captures natural conversational characteristics such as informal expressions, speaker interactions and context dependent humor making it suitable for real-world analysis. The main contributions of this work are as follows:

- We introduce a multimodal Hindi conversational dataset of 5203 utterances written in Devanagari script, containing natural code-mixed.
- We conduct a comprehensive study of humor detection using unimodal and multimodal approaches, including classical machine learning models, neural baselines, and advanced multimodal architectures.

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 33–39
October 29, 2026 ©2026 Association for Computational Linguistics

- We systematically compare early fusion, late fusion, and cross-modal representation learning approaches for humor detection.

2 Dataset

We collected data from the popular Indian web series *"Panchayat"* and manually annotated each utterance with humour and not humour labels, following prior multimodal humor dataset construction approaches (Hasan et al., 2019; Chauhan et al., 2021; Guo et al., 2024). The utterances are retained in their original Devanagari script, including naturally occurring code-mixed and transliterated English words without any normalization or transliteration. Each utterance having humour(1) in some kind of tone or visually was marked as humour else was marked not humour(0). Table 1 shows details about dataset.

Table 1: Dataset Statistics

Description	**Statistic**
Total Utterances	5203
Total Words	47289
Unique words	5733
Max Utterance Length	39
Min Utterance Length	1
Humour	1612
Not Humour	3591
Speakers	91
Max Video Length	14.56 s
Min Video Length	0.28 s
Total duration	8.91 Hours

3 Challenges

We initially employed an automatic speech recognition (ASR) system, Whisper for generating transcripts from the video data (Radford et al., 2023). While Whisper provides strong performance for multilingual speech recognition but it showed limitations in accurately transcribing Hindi conversational data in our dataset.

A significant challenge was the incorrect transcription of several Hindi words and code-mixed utterances. Many spoken words were either incorrectly recognized or replaced with semantically unrelated tokens, particularly in noisy or fast-paced dialogues.

Due to these limitations the automatically generated transcripts required extensive manual correction to ensure fidelity to the original spoken content. This manual intervention was necessary to preserve the authenticity of utterances and maintain consistency across the dataset.

These challenges highlight the difficulty of applying general purpose ASR systems to informal code-mixed Hindi conversational data.

4 Annotation Guidelines

Three experienced annotators were involved in labeling each utterance as Humour or Not Humour. Detailed annotation guidelines were provided to ensure consistency in understanding humour criteria and to reduce subjectivity in labeling. Prior to annotation, the annotators were trained on the labeling procedure and criteria for identifying humorous content. Utterances exhibiting sarcasm, situational comedy, exaggerated expressions or humorous delivery in tone and visual context were labeled as Humour while all other utterances were labeled Not Humour. To evaluate inter-annotator agreement, Fleiss' Kappa (Fleiss, 1971) was computed over all labeled utterances. The annotation process achieved a Fleiss' Kappa score of 0.78, indicating substantial agreement among annotators and ensuring the reliability of the dataset.

5 Baseline methods

To evaluate the effectiveness of the dataset we performed set of experiments using several baseline models ranging from traditional machine learning methods to advanced multimodal architectures.

5.1 Logistic Regression (LR)

Logistic Regression is used as a classical baseline for humor classification (Baltrušaitis et al., 2019). It learns a linear decision boundary to distinguish between humorous and non-humorous utterances. The model is trained on features extracted from individual modalities (text, audio, and visual) as well as their combinations to provide a simple yet effective benchmark.

5.2 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a feed-forward neural network used to model non-linear relationships in the input features. Compared to Logistic

Figure 1: Samples from the dataset.



Regression, MLP can capture more complex patterns across modalities. The model is trained on unimodal and multimodal feature representations to evaluate the benefit of non-linear feature learning.

5.3 Early Fusion

In early fusion, features from different modalities (Baltrušaitis et al., 2019) are concatenated at the input level to form a unified representation. This combined feature vector is then used to train a classifier. Early fusion enables joint learning across modalities but does not explicitly model interactions between them. Early Fusion uses learned per-modality projection layers with softmax-weighted combination, trained for 150 epochs with AdamW and cosine LR annealing.

5.4 Late Fusion

Late fusion combines predictions from individual modality-specific models (Rahman et al., 2020). Each modality is first processed independently, and their outputs are aggregated to produce the final prediction. This approach allows each modality to contribute separately while enabling flexible integration at the decision level. Unimodal rows in Table 3 are obtained from the same fusion architecture with non-target modalities masked to zero at the input, rather than separately trained single-modality models.

5.5 Cross-Modal Transformer (CMT)

The Cross-Modal Transformer (CMT) is a transformer-based architecture (Tsai et al., 2019) to capture interactions across different modalities using attention mechanisms. It models dependencies between text, audio, and visual features, allowing the system to learn richer multimodal representations for humor detection.

5.6 Disentangled Multimodal Representation Model (DMR)

The Disentangled Multimodal Representation Model (DMR) adapts the shared-private disentanglement architecture of MISA (Hazarika et al., 2020) by separating common and unique information across modalities. We omit the cross-modal similarity (invariance) loss between shared representations and use a simpler concatenation-based fusion in place of a transformer fusion layer. We retain the descriptive name DMR to reflect this simplified task-specific adaptation.

6 Features Extraction

For textual feature extraction, we use the pre-trained MuRIL (Khanuja et al., 2021) to obtain contextual embeddings for each utterance, where the [CLS] token representation is used as the sentence-level feature. For acoustic features, we employ a pre-trained HuBERT (Hsu et al., 2021), which extracts frame-level representations that are aggre-

gated using mean and max pooling to obtain fixed-length embeddings. For visual features, we use a pre-trained VideoMAE (Tong et al., 2022) model, where uniformly sampled frames from each video clip are processed to generate spatio-temporal representations, followed by mean pooling. All features are L2-normalized to ensure consistency across modalities.

7 Experiments and analysis

We evaluate the effectiveness of the dataset using multiple baseline models across different modality combinations. All experiments are conducted using 5 fold cros validation to ensure robustness and reduce bias due to data splits.

The feature representations for text, audio, and visual modalities are extracted using pre-trained models and standardized before training. For multimodal experiments, different combinations of modalities (T, A, V) are evaluated, including unimodal, bimodal, and trimodal settings. Class imbalance in the dataset is handled using class-weighted loss. We report Accuracy (A), Precision (P), Recall (R), and F1-score (F1) as evaluation metrics for humor classification.

7.1 Results and Analysis

We evaluate all models under different input configurations including unimodal (T, A, V), bimodal (T+A, T+V, A+V), and trimodal (T+A+V).

For LR and MLP, MLP consistently performs better than LR across most settings. In unimodal cases, MLP achieves an accuracy of 67.23% (T), 61.62% (A), and 59.49% (V), whereas LR achieves 61.37% (A+V) as its best unimodal/bimodal performance. In the trimodal setting, LR achieves 61.35% accuracy and 59.86% F1-score, while MLP achieves higher values with 65.19% accuracy and 60.32% F1-score. We show the results in Table 2

Table 2: Results of Logistic Regression and Multilayer Perceptron

Labels	LR				MLP			
	A	P	R	F1	A	P	R	F1
T	56.91	57.46	58.70	55.51	**67.23**	60.19	58.14	54.46
A	60.18	60.37	62.11	58.75	61.62	57.42	58.00	57.10
V	58.12	56.25	57.23	55.52	59.49	56.39	56.92	55.52
T+V	58.47	56.69	57.74	55.93	63.56	**60.83**	60.63	58.61
T+A	60.22	60.45	62.20	58.80	65.00	59.89	60.06	59.50
A+V	**61.37**	61.22	63.11	59.85	62.02	58.37	59.31	58.31
T+A+V	61.35	**61.26**	**63.16**	**59.86**	65.19	60.69	**60.93**	**60.32**

Table 3: Results for Early and Late fusion

Labels	Early fusion				Late fusion			
	A	P	R	F1	A	P	R	F1
T	62.58	63.09	**65.23**	61.31	63.02	**63.65**	**65.95**	61.87
A	63.18	58.41	58.98	58.54	63.69	58.92	59.54	59.09
V	59.83	57.34	58.41	56.90	60.08	57.45	58.52	57.10
T+V	64.85	63.12	65.18	62.71	64.54	62.84	64.88	**62.41**
T+A	65.65	62.08	63.43	62.25	64.79	61.26	62.58	61.40
A+V	64.23	60.20	61.22	60.35	64.71	60.08	60.81	60.29
T+A+V	**66.90**	**63.18**	64.55	**63.45**	**65.56**	62.16	63.61	62.32

Table 3 shows Early Fusion achieves 66.90% accuracy and 63.45% F1-score in the T+A+V setting. Late Fusion achieves slightly lower performance with 65.56% accuracy and 62.32% F1-score in the same configuration. In unimodal settings, Early Fusion achieves 62.58% (T), 63.18% (A), and 59.83% (V) accuracy, while Late Fusion shows 63.02% (T), 63.69% (A), and 60.08% (V) accuracy.

Table 4 results show CMT achieves its best performance in the trimodal setting with 67.69% accuracy and 68.06% F1-score. DMR achieves the highest overall results with 69.04% accuracy and 69.11% F1-score in the T+A+V configuration.

Table 4: Results for Cross-Modal Transformer and Disentangled Multimodal Representation Model

Labels	CMT				DMR			
	A	P	R	F1	A	P	R	F1
T	66.73	**68.92**	66.73	67.44	66.10	**69.76**	66.10	67.16
A	61.62	64.89	61.62	62.73	62.41	65.68	62.41	63.51
V	57.52	65.00	57.52	59.04	61.18	63.70	61.18	62.10
T+V	67.44	67.45	67.44	67.35	68.23	69.37	68.23	68.68
T+A	**67.86**	68.26	**67.86**	67.87	68.19	68.99	68.19	68.53
A+V	64.25	65.83	64.25	64.71	64.94	65.80	64.94	65.32
T+A+V	67.69	68.84	67.69	**68.06**	**69.04**	69.27	**69.04**	**69.11**

Table 5: Some correct and incorrect predicted samples.

No.	Hindi Utterances	Actual	Predicted	
			Unimodal (T)	Multimodal (T+V+A)
1	खाना वाना खा लिया आराम करते हैं यार	Not Humour	Humour	Not Humour
2	अब एट्टी टू से नाइन्टी नाइन तो पहुँचने से रहा इस जिंदगी में	Humour	Not Humour	Humour
3	ठीक है लेकिन बड़का ऊँगली तो आत्माराम के लिए ही रहेगा	Humour	Not Humour	Humour
4	यू शुड लव व्हाट यू डूइंग	Not Humour	Not Humour	Humour
5	चलिये सब लोग खाना तैयार है	Not Humour	Not Humour	Humour
6	और गलती से हमने सड़क बना दिया तो उसकी छाती पे साँप लोट जाएगा देख लीजिएगा	Humour	Humour	Not Humour

7.2 Overall Trimodel Comparison

We present the overall model comparison under the trimodal (T+V+A) setting in Table 6 to evaluate the effectiveness of different approaches. The results show that baseline methods such as LR and MLP achieve comparatively lower performance while fusion-based models provide consistent improvements. DMR achieves the best overall performance, demonstrating a strong balance between precision and recall.

Table 6: Overall trimodel Comparison

Model	Acc	Prec	Rec	F1
LR	61.35	61.26	63.16	59.86
MLP	65.19	60.69	60.93	60.32
Early Fusion	66.90	63.18	64.55	63.45
Late Fusion	65.56	62.16	63.61	62.32
CMT	67.69	68.84	67.69	68.06
DMR	**69.04**	**69.27**	**69.04**	**69.11**

8 Error Analysis

In this section we take and compare predictions of samples from the dataset. We take some samples that were predicted incorrectly by model to analyze models strength and weakness in predicting a sample.

Table 5 shows some samples from dataset which were predicted by the Unimodal and multimodal. Comparision of Unimodal(text) is being done here with multimodal to see where it fails to predict Humour and Not Humour utterance. For understanding the context of dialogues they are translated in English (the dataset does not contains), for example in Table 5 utterance number 1 ''खाना वाना खा लिया आराम करते हैं यार'' (Had food and all, let's relax now, man.), here Abhishek is talking to his friend Siddharth in relax tone with not much facial expressions. The correct label is Not Humour but the unimodal(T) predicts this sample as Humour whereas the multimodal predicts it correctly as Not Humour because of the tone and visual. The neutral tone and visual frame helps multimodal to determine this correctly as Not Humour, visual frames shown on figure 2 shows expressions of speaker(Abhishek).

Figure 2: Utterance number 1 visual frames.



Also in utterance number 3 in Table 5 Pushpa says ''ठीक है लेकिन बड़का ऊँगली तो आत्माराम के लिए ही रहेगा'' (Okay, but the big finger will remain for Atmaram only.) while saying this the facial visuals in figure 3 can be seen as sarcastic and smirk like face made while saying it, the multimodal catches this correctly and predicts it correctly as Humour unlike unimodal which predicted it Not Humour.

Figure 3: Utterance number 3 visual frames.



While the multimodal predicted many samples correctly it also struggled and failed to predict some of the correct labels as shown in Table 5. Utterance 5 speaker(Rinki) says ''चलिये सब लोग खाना तैयार है'' (Come on everyone, the food is ready.), unimodal fails to predict the correct label here because of the joyful and sarcastic facial features that can be seen in fig 4, expressions of Rinki seems sarcastic but was a normal line said by her.

Figure 4: Utterance number 5 visual frames.



Another wrong example from Table 5 is utterance number 6 where speaker(Brij Bhushan) says ''और गलती से हमने सड़क बना दिया तो उसकी छाती पे साँप लोट जाएगा देख लीजिएगा'' (And if by mistake we build the road you will see that a snake will crawl on his chest.), the actual label for this utterance is Humour but the multimodal fails to predict it correctly because of the hyper, sad and angry tone and visual expressions of Brij Bhushan. Figure 5 shows visual frames and expression.

Figure 5: Utterance number 6 visual frames.



9 Conclusion

In this work, we presented a multimodal dataset for humour detection in Hindi conversational data derived from the web series *"Panchayat"* (8.91 hours in length). The dataset preserves utterances in their original Devanagari script as spoken, including code-mixed and transliterated English words without any normalization or translation ensuring natural conversational characteristics.

We evaluated multiple baseline models across unimodal, bimodal, and trimodal settings. Results consistently show that multimodal approaches outperform unimodal ones, highlighting the importance of integrating textual, acoustic, and visual cues for effective humour recognition. Among the evaluated methods, advanced multimodal models achieve the best performance, demonstrating the benefit of modeling cross-modal interactions.

We plan to extend the HindiHumour dataset with multi label annotations to capture the fine-grained and multifaceted nature of conversational humour. Each utterance with one or more humour categories including sarcasm, situational comedy, irony, exaggeration, self-deprecation, etc. enabling multi-task and multi-label learning approaches.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, San Diego, California. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-Philippe Morency, and Soujanya Poria. 2021. M2H2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 773–777.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Hongyu Guo, Wenbo Shang, Xueyao Zhang, Shubo Zhang, Xu Han, and Binyang Li. 2024. MUCH: A multimodal corpus construction for conversational humor recognition based on chinese sitcom. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11692–11698, Torino, Italy. ELRA and ICCL.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Antonios Kalloniatis and Panagiotis Adamidis. 2025. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58:43.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Chong Min Lee, Su-Youn Yoon, and Lei Chen. 2016. Can we make computers laugh at talks? In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 173–181, Osaka, Japan. The COLING 2016 Organizing Committee.

Amruta Purandare and Diane Litman. 2006. Humor: Prosody analysis and automatic recognition for F*R*I*E*N*D*S*. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215, Sydney, Australia. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in Neural Information Processing Systems*, 35:10078–10093.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Heads That Ask: Prosody of Questionhood in Whispering and Mutual (In)visibility of Speakers

Nasim Mahdinazhad Sardhaei
Humboldt-Universität / Germany
nasimmahdinazhad@gmail.com

Marzena Żygis
Leibniz-ZAS / Germany
Humboldt-Universität / Germany
zygis@leibniz-zas.de

Hamid Sharifzadeh
Unitec Institute of Technology / New Zealand
hsharifzadeh@unitec.ac.nz

Abstract

Whispering differs from voiced speech in the lack of fundamental frequency (F0) and is characterized by substantially reduced amplitude. This production study investigates how head movements are modulated in whispering when speakers are visible or not, while producing polar questions with rising intonation versus statements with flat intonation. An experiment with 15 German speakers shows that head movements are more pronounced in questions than in statements, with the strongest effects observed in normal questions under mutual visibility. The findings suggest head movements serve as a flexible multimodal prosodic resource, aligning with acoustic cues in normal speech under favorable communicative conditions while functioning independently as compensatory mechanism in whispered speech.

1 Introduction

Communication extends beyond spoken language, involving both auditory and visual channels. While speaking, we synchronize head, face, hand, and body movements with our speech, fulfilling communicative functions. These movements are closely linked to acoustic prosody cues, emphasizing key information, marking rhythmic pauses, or signaling sentence types intonation, as e.g. in questions versus statements (Cavé et al., 1996; Swerts and Krahmer, 2004; House, 2002; Dohen and Loevenbruck, 2008).

Among these various gestures, orofacial expressions and head movements such as eyebrow raises, nods, and head tilts, have been shown to carry prosodic information, complementing acoustic cues like F0. Research has consistently found significant relation between F0 and both orofacial and head movements. For example, changes in eyebrow position have been linked to variations in F0 (Cavé et al., 1996; Guaïtella et al., 2009), while rigid head motions also reflect prosodic patterns (Ishi et al., 2007; Yehia et al., 2002). Importantly, listeners can distinguish echoic questions from declarative statements based solely on visual head dynamics, even in simplified animations that preserved only rigid head motion indicating that head-motion cues alone can signal prosodic structure (Cvejic et al., 2010). In European Portuguese, a downward head movement (head fall) characterized yes–no questions, whereas statements displayed subtler pitch-like head motions (Cruz et al., 2015).

However, speech communication may occur in different settings, where speakers may adjust both verbal and non-verbal behavior to contextual constraints, affecting gesture production. One key factor in this interplay between speech and gesture is mutual visibility, a physical parameter that can impose varying communicative constraints, affecting speakers' strategies to convey their intended meaning effectively. Depending on whether the listener can see the speaker, the latter must adapt their verbal and non-verbal signals to the constraints of the physical medium (Bavelas et al., 2008; De Ruiter et al., 2012; Żygis and Fuchs, 2023).

The role of speakers' mutual visibility in gesture production has been subject of debate. Some studies argue that gestures primarily serve the listener, leading to fewer gestures when the listener is not visible (Bavelas et al., 1992; Krauss et al., 1995; Alibali et al., 2001; Emmorey and Casey, 2001). In contrast, others (Rimé, 1982; De Ruiter, 1998; Mol et al., 2011) propose that gestures facilitate the speaker's own speech production processes rather than serving purely communicative purposes, in which case visibility should have little effect. Notably, these studies have predominantly focused on voiced speech and specific types of gestures such as pointing. The present study manipulates speaker (in)visibility to examine how speech and head movements interact in whispered speech mode, where the lack of fundamental frequency

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 40–49
October 29, 2026 ©2026 Association for Computational Linguistics

(F0) imposes an extra communicative constraint. In this mode, altered phonation produces voiceless acoustic signals that can challenge speech intelligibility.

Some earlier studies have examined the interaction between acoustic signals and orofacial gestures in whispered speech, with no attention to other body gestures such as head movements. For instance, Dohen and Loevenbruck (2008) reported that perception of focus in whispered French improved when auditory and visual cues such as eyebrow movements were combined. Tao and Busso (2014) showed that lip movements enhanced the accuracy of recognizing whisper. Żygis and Fuchs (2023) revealed higher eyebrow raising in whispered German questions. Similarly, Mahdinazhad Sardhaei et al. (2022) showed that in Farsi, facial expressions and word duration adjusted in whisper to offset the absence of F0, while Mahdinazhad Sardhaei et al. (2024) further showed that longer word durations and specific orofacial expressions acted as compensatory mechanisms in whisper to convey the intonation of a given sentence type. Building on this foundation, our study shifts the focus from orofacial expressions to head movements, investigating whether they contribute to intonational distinctions between yes–no questions and statements in whisper, where F0 is not available.

This study will examine the interplay between head movements and acoustic cues of prosody across sentence types under the above-mentioned constraints through the lens of two alternative hypotheses. The first, the trade-off hypothesis (Bangerter, 2004; Melinger and Levelt, 2004; Van der Sluis and Krahmer, 2004; De Ruiter, 2006; Van der Sluis and Krahmer, 2007), suggests that speakers tend to use gestures to ease communicative burden when verbal expression is difficult, but rely on speech when gesturing becomes demanding. The second, hand-in-hand hypothesis, proposes that gestures are produced not solely for communicative purposes but for the speaker's own cognitive or articulatory benefits. According to this view, gestures go hand in hand with speech (Goldin-Meadow, 2009; So et al., 2009). If the trade-off hypothesis holds true, we might expect head movements to compensate for the lack or distortion of F0. While, a positive relationship between head movements and acoustic parameters such as duration and mean amplitude would provide tentative support for the hand-in-hand hypothesis.

To summarize, this study aims to examine the composite (i.e., combined measures across pitch, yaw, and roll axes) velocity and magnitude of head movements, as well as their total extent of movement. Our hypotheses are:

1. Different measures of head movement will be greater in whispered questions than normally spoken questions while speakers do not see each other.

2. According to the trading hypotheses, we explore the following relations:

a. If the trade-off hypothesis holds true, we might observe a negative correlation between acoustic parameters (amplitude and sentence final word duration) and different head movements.

b. Alternatively, if the hand-in-hand hypothesis holds true, a positive correlation between acoustic parameters and head movements might emerge.

2 Methods

2.1 Participants

Fifteen native German speakers (4 male, 11 female; mean age = 33.15, SD = 12.60) with no motor, hearing, or language impairments participated. All had normal or corrected vision, provided informed consent, and received monetary compensation.

2.2 Stimuli

The study used 40 German sentences, consisting of 20 structurally identical statement–question pairs differing only in punctuation (question mark vs. period). Target words appeared in sentence-final position and were bisyllabic content words with initial syllable stress. Each initial syllable conformed to a CVC structure, beginning with a bilabial stop (/p/, /b/, or /m/) followed by unrounded vowels (/a/, /ɛ/, or /ɪ/). Vowels differed in height, with /a/ showing the widest lip aperture and (/ɪ/) the smallest. The target words were in sentence-final position, where intonation rises in (polar) questions and remains flat or falls in statements (for German, see Grice et al., 2005). The experiment had two contextual conditions: speech mode (normal and whisper) and visibility (visible vs. invisible). In the visible condition, participants and the confederate faced each other; in the invisible condition, an artificial wall separated them. As the result, four stimulus blocks were designed, each containing sentences classified as questions or statements:

a. Normal speech: Speakers see each other, speakers do not see each other.

b. Whisper speech: Speakers see each other, speakers do not see each other.

To minimize order effects, both block presentation and sentence sequence within blocks were randomized for each speaker. Each speaker produced 160 sentences in total (40 sentences × 2 visibility conditions × 2 speech modes).

2.3 Data collection

The experiment was conducted in a soundproof lab with video (Canon EOS1300D, 30 Hz) and audio (Sennheiser ME62, 44,100 Hz) recordings. Participants rehearsed intonations before the session, with a confederate. In the experiment, the confederate produced either a question or a statement and the informant responded with the opposite sentence type intonation; questions followed statements and vice versa. Thus, only intonation was modified while sentence content and word order remained unchanged. For example:

• Confederate: Er mag diese Piste. ("He likes this slope.")

• Informant: Er mag diese Piste? ("He likes this slope?")

To ensure accurate speech mode production and proper prosody, the confederate (a trained native German speaker who remained the same across all sessions) practiced a few sentences with each participant beforehand. Throughout the experiment, this interlocutor monitored participants' productions, and asked for repetitions when whisper quality or intonation was unsatisfactory. After data collection, a second native German speaker reviewed all utterances to verify intonational accuracy.

2.4 Data preparation

To align the timing of video and audio files of varying lengths, we used Adobe Premiere Pro 24.4 for synchronization. Audio recordings were annotated in PRAAT 6.4.20 (Boersma and Weenink, 2019), focusing on sentence-final words. Start and end times from the PRAAT TextGrid were extracted and used in a Python script to automatically segment and trim the corresponding video files. Using PRAAT 6.4.20, we manually labeled the boundaries of the sentence-final word (onset/offset), stressed syllable (onset of initial stop to syllable-final segment), and unstressed syllable (onset/offset). Our measurements included (a) sentence-final word duration and (b) amplitude difference (mean amplitude of the word-final minus word-prefinal syllable). Calculating amplitude difference provided a normalized estimate of amplitude while partially reducing variability due to speaker–microphone distance, although it does not fully eliminate all sources of variation.

2.5 Measurement of head movement

To quantify head movement, we used OpenFace 2.0 (Baltrušaitis et al., 2018) to track facial structures. Videos of target words were processed at 29.969 FPS. OpenFace can track head movements in radians per frame. Following OpenFace conventions, we extracted head pitch (Rx, up-down nodding), yaw (Ry, left–right rotation), and roll (Rz, side-to-side tilt) to quantify head movements. Our procedures for deriving the target measures of head movements are as follows:

Velocity of head pitch, yaw, roll: The velocity of head movements represents the rate of change in head pose angles between consecutive frames, measured per unit of time. To assess head movement velocity over an entire word, we averaged the measures to summarize movement for each target word over N frames. The mean velocity over a word, considering N frames, was calculated as follows:

$$\frac{1}{N-1}\sum_{t=2}^{N}|\text{Velocity}(t)| \tag{1}$$

Magnitude of head pitch, yaw, roll: To this end, we defined a neutral baseline (nb) by averaging of the first 3 frames. Over an entire word, we calculated the mean absolute magnitude as:

$$\frac{1}{N-3}\sum_{t=4}^{N}|\text{PitchRx}(t)-\text{PitchRx}(nb)| \tag{2}$$

We also calculated composite measures which captured overall movement by combining pitch, yaw, and roll using Euclidean norms for velocity and magnitude, as follows:

Composite velocity: it gives a scalar measure of how fast the head moves overall and was calculated as:

$$\frac{1}{N}\sum_{t=1}^{N}\sqrt{(VR_x(t))^2+(VR_y(t))^2+(VR_z(t))^2} \tag{3}$$

Composite movement magnitude: it was computed per frame as the Euclidean norm of pitch,

yaw, and roll components and then averaged across N frames:

$$\frac{1}{N}\sum_{t=1}^{N}\sqrt{(IR_x(t))^2+(IR_y(t))^2+(IR_z(t))^2} \tag{4}$$

where IR_x, IR_y, and IR_z denote pitch, yaw, and roll magnitudes at frame t.

Total head movement: it was quantified as a direction-independent measure by summing movement across all axes during the word:

$$\sum_{t=2}^{N}\sqrt{(\Delta R_x(t))^2+(\Delta R_y(t))^2+(\Delta R_z(t))^2} \tag{5}$$

2.6 Statistics

We conducted statistical analyses in R Studio 4.2.3. using the lme4 (Bates et al., 2020) and emmeans packages (Lenth et al., 2018). Linear mixed models assessed the effects of Speech Mode (normal vs. whisper), Visibility (visible, invisible), Sentence Type (question, statement) on head movements. Reference levels were 'normal speech', 'visible', and 'statement'. To investigate whether the dependent variables (Composite velocity, magnitude, total movement, as well as velocity and magnitude of head pitch, yaw, roll) exhibited the most pronounced effects under marked conditions, i.e. whispered speech, questions, and invisibility of speakers, a three-way interaction of Speech Mode, Visibility, and Sentence Type was incorporated into the models. The models accounted for random effects by including random intercepts for Word and Participant, along with by-Participant and by-Word random slopes for Speech Mode, Visibility, Sentence Type, and their interactions. Some initial models failed to converge and/or produced singular-fit warnings. To obtain parsimonious and identifiable random-effects structures, we followed two simplification procedures. First, random-effects terms involved in perfect correlations (i.e., correlations of ±1) were removed, following the recommendations of (Winter, 2020). Second, when singularity persisted, we inspected the random-effects structure using the rePCA() function in the lme4 package. Following the rationale proposed by (Bates et al., 2015), the function rePCA() evaluates the effective dimensionality of the random-effects structure by performing a Principal Components Analysis (PCA) of the estimated random-effects variance–covariance matrix. The random-effects structure is decomposed into orthogonal components, and components with very small or zero standard deviations indicate dimensions that are not supported by the data and therefore contribute to model singularity. The output also provides a rotation matrix that maps the original random-effects parameters onto the orthogonal principal components, allowing identification of redundant variance components. These components were removed through simplification of the random-effects structure (e.g., by eliminating the corresponding random slopes), resulting in a more parsimonious model. Our final model used for all dependent variables was:

```
lmer(duration ~
Speech_mode*Visibility*sentence_type
+ (1 + Speech_mode + Visibility +
    sentence_type | participant)
+ (1 | word), data = German_data1)
```

The models were also checked with respect to (i) the normality of the residuals, (ii) their homoscedasticity and (iii) multicollinearity among the factors. To assess the significance of fixed effects in the final model, we conducted an Analysis of Variance (ANOVA) using Type III sum of squares. This approach tests each predictor's unique contribution to the model while accounting for all other variables and their interactions. The Satterthwaite approximation was applied to estimate degrees of freedom, providing more accurate p-values in the presence of random effects. This method is particularly appropriate given the mixed-effects structure of our model and the unbalanced nature of the dataset. Finally, we then used the emmeans() function with Tukey adjustment to conduct pairwise comparisons of mean values for the dependent variables across different conditions.

Additionally, we used lmer() to examine the relationship between composite velocity and composite magnitude of head movements and two acoustic parameters of word duration and amplitude difference. We assumed that these movements may compensate for the absence of F0 in whispering. Accordingly, under the trade-off hypothesis, we would expect a negative relationship, whereas under the hand-in-hand hypothesis, we predicted a positive relationship. Continuous variables were standardized (z-scored) per speaker, and by-speaker and by-word random intercepts were included in the

models. It should be pointed out that we dispensed with a classic correlation modelling as our data included repeated (dependent) data points. An example is provided below:

```
lmer(composite_velocity ~
word_duration +
(1|participant) + (1|word))
```

3 Results

Here, we first report the results for the composite measures, and total movement, followed by a breakdown of the relative contributions of head pitch, yaw, and roll to the composite outcomes. Finally, we examine the extent to which different head movements and acoustic parameters exhibit a linear relationship.

3.1 Composite velocity

The mixed effect model revealed a significant interaction between Speech mode *Visibility*Sentence type ($F(1, 2234.02) = 9.959$, $p<.01$). As Figure 1 shows, head composite velocity was the greatest when speakers produced questions in the normal visible condition. Pairwise comparisons revealed greater velocity in visible normal questions than in normally produced statements under visible ($t = 3.69$, $p = .03$) and invisible ($t = 3.40$, $p = .025$) conditions.

3.2 Composite magnitude

The mixed effect model for head composite movement magnitude showed a significant interaction between Speech mode*Visibility*Sentence type ($F(1, 2217.06) = 4.125$, $p<.05$). Pairwise comparisons showed that head movement was greater for normally produced questions under visible condition than for statements produced in both visible ($t= -6.61$, $p <.001$) and invisible ($t= -5.78$, $p <.001$) conditions. Also, greater head magnitude was observed for visible normal questions compared to invisible whispered statements ($t= -3.53$, $p = 0.046$, Figure 1).

3.3 Total movement

For total head movement, the interaction between Speech mode*Sentence type turned out to be significant ($F(1, 2231.89) = 26.705$, $p<.001$). As Figure 2 (right) shows, normally produced questions had the largest total movement. Pairwise comparisons indicated that normally produced questions had a larger total movement than normal statements($t= -6.23$, $p <.001$). on the other hand, whispered questions showed a higher total movement than whispered ($t= -3.12$, $p = 0.01$) and normal ($t= -3.02$, $p = 0.03$) statements.

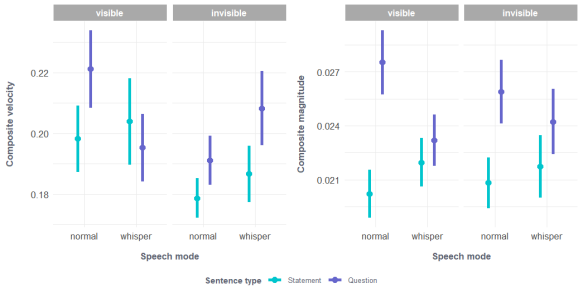


Figure 1: Composite velocity (left) and magnitude (right) of head movement in the sentence-final word: interaction between speech mode, (in)visibility, and sentence type.



Figure 2: Total head movement showing an interaction between speech mode, and sentence type.

3.4 Contribution of individual axes to composite head movement

To clarify which axis of head movement primarily drove the composite results, we decomposed the composite measures into their individual pitch (Rx), yaw (Ry), and roll (Rz) components. Among these, head pitch emerged as the dominant contributor. Both pitch velocity and magnitude yielded significant interactions between Speech mode*Visibility*Sentence type (velocity: $F(1, 2233.54) = 10.916$, $p<.001$; and magnitude: $F(1, 2201.21) = 4.629$, $p<.05$; see Figures 3 and 4). Specifically, the head pitch velocity was the greatest when speakers produced questions in selected conditions (visible normal and whisper invisible). However, pairwise conditions revealed only few significant differences with greater pitch velocity in normal questions under visible condition than normally produced statements under visible ($t = 3.74$, $p = 0.013$) and invisible ($t= 3.96$, $p = 0.017$) con-

ditions. The close correspondence between pitch and composite patterns suggests that head nodding accounts for the largest share of variance in overall head displacement.

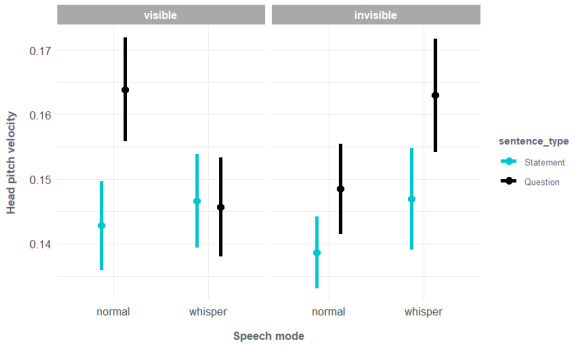


Figure 3: Head pitch velocity in the sentence-final word: three-way interaction between speech.

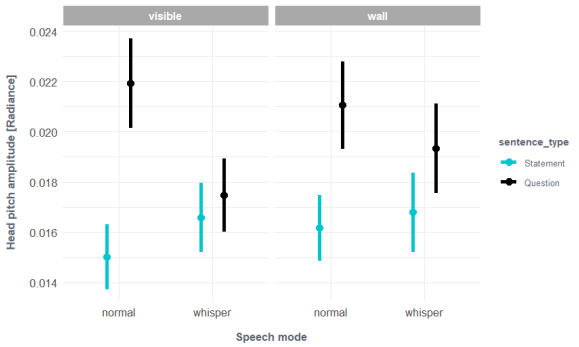


Figure 4: Head pitch magnitude in the sentence-final word: three-way interaction between speech mode, visibility, and sentence type.

By contrast, head roll movement did not show any significant effects. Only, yaw magnitude yielded a significant two-way interaction between speech mode and sentence type ($F(1, 2213.35) = 4.914$, $p<.05$), with questions produced with greater yaw, and a larger difference between questions and statements in normal speech than in whisper ($t= 2.87$, $p = 0.03$).

3.5 Relationship between acoustic cues and head movements

At this stage, we analyzed the relationship between head movements (composite velocity and composite magnitude) and two acoustic parameters including word duration and mean amplitude difference. All continuous variables were z-scored for each speaker. Although the observed relationships were small, some were statistically significant.

The analysis revealed a negative relationship between word duration and head composite velocity in both normal ($t = -3.22$, $p <.001$) and whisper ($t = -2.53$, $p = 0.011$) modes, indicating faster movement for shorter words (see Figure 5). In contrast, composite head movement magnitude was positively correlated to word duration only in whispered speech ($t = 2.82$, $p < .01$), with longer words showing larger movements.

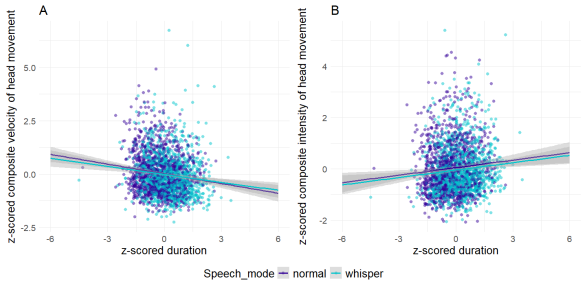


Figure 5: Scatterplots for sentence-final word duration in relation to head movement composite velocity and magnitude.

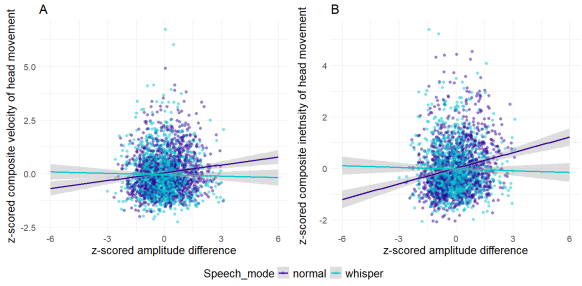


Figure 6: Scatterplot for amplitude difference in relation to head movement composite velocity and magnitude.

Mean amplitude difference and head composite velocity were positively correlated in normal speech mode ($t= 4.70$, $p<.0001$, see Figure 6) suggesting louder words were associated with faster head movement. No significant result was observed for whispered speech. Similar results were found for composite head movement magnitude, with a positive relationship to mean amplitude difference only in normal speech ($t= 5.23$, $p<.0001$).

4 Discussion

To our knowledge, this is the first study to examine head movements in conveying sentence-type intonation under communicative constraints, particularly in whispered speech, while manipulating interlocutor mutual visibility. The findings are discussed in relation to the trade-off and hand-in-hand hypotheses and the broader literature on multimodal prosody. We focused on sentence-final

words where intonational contrasts between polar questions and statements are primarily realized. Even within this brief part of the sentence, significant differences emerged across all head movement measures with faster and larger head movements in questions than in statements. This pattern was evident in all head movement measures. Consistent with earlier research (e.g. Cvejic et al., 2010; Yehia et al., 2002), these results confirm that rigid head motion carries prosodic information and contribute to the signaling of sentence type. Questions, typically characterized by rising intonation, appear to involve enhanced multimodal prosodic marking compared to statements. The ability of listeners to distinguish sentence types based solely on head motion (Cvejic et al., 2010) further supports the communicative relevance of the patterns observed in our study.

Our initial prediction was that the strongest head gestures would occur under the most communicatively constrained condition; that is, whispered questions produced in the absence of speakers' mutual visibility. However, the results did not confirm this hypothesis. Instead, the most pronounced head movements were consistently found in the least constrained condition, i.e., normally produced questions in the visible condition with the largest effects across composite velocity, composite magnitude, and total movement. Some previous studies have also reported increased gesture production when speaker and listener can see each other than when there is a visual barrier between the two (e.g. Alibali et al., 2001; Mol et al., 2011; Barker and Kita, 2026). Boosting of head movements in normally produced questions when interlocutors are visible may reflect a general prosodic marking strategy that is amplified when full communicative resources are available. Under these circumstances, head movements appear to function in parallel to acoustic prosodic cues such as F0, reinforcing the sentence type contrast through visual signaling. This redundancy may reflect an audience-oriented strategy (Bavelas et al., 1992), whereby speakers enhance communicative clarity when resources are optimal.

The role of visibility, however, proved selective rather than uniformly amplifying. Although a significant interaction between speech mode, visibility, and sentence type indicates that visibility contributes to gestural behavior, this effect is not consistent across all measures as total head movement was unaffected by (in)visibility. Thus, visibility may not globally influence head gesture production but selectively modulate specific kinematic parameters.

To better interpret the findings, we examined the relative contributions of individual head movement axes revealing head pitch as the primary vehicle of prosodic head movement in our data. Head pitch emerged as the most prominent marker of sentence type contrasts, showing increased velocity and magnitude for normally produced questions, when the interlocutor was visible while head yaw and roll playing very small roles. This close correspondence between head pitch patterns and the composite measure indicated that head nodding accounts for the largest share of variance in overall head displacement, and that the sentence-type contrasts observed in the composite measures are primarily driven by up-and-down head rotation. Another point to notice was significant magnitude of head yaw movement and its contribution to visually marking sentence-type distinctions. The co-occurrence of pitch and yaw magnitude in questions suggests that these two axes may not operate entirely independently. Instead, speakers may use coordinated gesture combining up–down (pitch) and left–right (yaw) head motions. In other words, a speaker may nod while slightly turning the head. The resulting movement is not purely nodding or turning, but a smooth trajectory that has incorporates both components. Such multi-axis coordination in prosodic gesturing has been noted in previous accounts of head movement behavior (Ishi et al., 2007),(Yehia et al., 2002), and may reflect a naturally integrated gestural strategy for marking sentence-type boundaries visually.

Our results can be interpreted considering two hypotheses: the trade-off hypothesis which posits that gestures are produced to support listener comprehension (Alibali et al., 2001), and the hand-in-hand hypothesis according to which gesturing helps the speaker reduce cognitive load (Goldin-Meadow, 2009).

In this study, the relationship between head movements and acoustic parameters (word duration and mean amplitude difference) does not straightforwardly support either hypothesis alone. Instead, it points to a flexible system in which head movements reflect different mechanisms depending on speech mode, sentence type, and speaker visibility, with patterns consistent with both accounts.

For sentence-final word duration, shorter words were accompanied by faster composite velocity in

both normal and whispered speech which suggests a timing-based coordination mechanism: head movement dynamics are adjusted to align with the temporal structure of speech rather than reflecting a simple scaling predicted by either hypothesis. On the other hand, in whispered speech, longer words were associated with greater composite magnitude which may be consistent with the hand-in-hand hypothesis. However, rather than drawing a definitive conclusion, we interpret this parallel enhancement of acoustic information and head movements as a compensatory strategy for the absence of F0 and degraded amplitude where speakers may rely on alternative prosodic cues such as longer duration coupled with more pronounced head movements to convey intended intonation. Together, these patterns indicate that in whispered speech, speakers flexibly adjust their head movements in tandem with word duration balancing temporal constraints with the need to compensate for reduced acoustic prosodic cues.

The amplitude difference analysis revealed louder sentence-final words were associated with faster and larger head movements. The synchronous strengthening of acoustic and gestural signals supports the hand-in-hand hypothesis (Goldin-Meadow, 2009) viewing gesture and speech as parallel systems driven by the same underlying demands. In normal speech, speakers may coordinate acoustic and visual cues of prosody either to enhance listeners' perception of prosodic contrasts or increase communicative effectiveness of their own utterance (Van der Sluis, 2005; Van der Sluis and Krahmer, 2007). Moreover, this finding resonates with the motor–prosodic coupling hypothesis (McNeill, 1992), which holds that prosodic prominence engages greater biomechanical energy across both speech articulators and the body. Louder speech requiring greater articulatory effort may co-occur with enhanced head and upper-body movement activity (Hadar et al., 1983; Munhall et al., 2004). In this view, greater movement magnitude reflects a coordinated motor response to prosodic emphasis rather than a direct consequence of acoustic intensity alone. No such relationship was found in whispered speech, likely because amplitude differences are much degraded in this speech mode.

5 Conclusion

In summary, by examining the relationship between the acoustic signal and oro-facial expressions when speakers (i) speak normally or whisper, and (ii) do or do not see each other, we found the overall advantage of visible, normally spoken questions over all other conditions. The results are interpreted in terms of two hypotheses, i.e., the hand-in-hand and trade-off hypotheses. The relationship between acoustic properties and head gestures proved to be mode-dependent rather than uniform. In normal speech, the positive coupling between amplitude difference and both composite velocity and magnitude points to a hand-in-hand mechanism, whereby acoustic and head gesture are co-produced and mutually reinforcing. In whispered speech, however, this coupling breaks down, and head movements appear to operate more independently: longer words are accompanied by greater head movement magnitude, while shorter words trigger faster gestural velocity suggesting a compensatory mechanism to preserve prosodic contrasts in the absence of F0, while also balancing temporal constraints by faster movement. Although head movements may require more effort than other orofacial expressions, the observed patterns indicate that speakers use their heads as an important visual resource, reinforcing prosodic marking across different communicative conditions. Nevertheless, previous research has also shown that orofacial movements serve as important cues for marking sentence-type intonation in whisper. Further perceptual studies could investigate the relative contributions of head movements and orofacial expressions in distinguishing different prosodic patterns in whisper vs. normal speech. In particular, such studies could determine whether head movements play a compensatory role in whispering by helping listeners recover prosodic distinctions when F0 is absent. The present analyses focused on direction-independent composite measures capturing how large and how fast speakers moved their heads to mark prosodic contrasts, without distinguishing the specific orientations of those movements. While this approach provides a summary of prosodic head gesturing, future studies incorporating head movement trajectories would provide a more detailed account of the directional and temporal dynamics of prosodic head gesturing. Such work would benefit further from being conducted in naturalistic communicative settings with spontaneous interactions to capture a more ecologically valid picture of how head movements signal sentence-type intonation in everyday, particularly whispered speech. Our findings are based on a relatively small sample of 15 participants, which

may constrain the generalizability of the results. This limitation highlights the need for further research involving larger and more diverse participant samples. Finally, it remains an open question whether the patterns observed in German generalize to other languages with different prosodic systems, or to sentence types beyond the polar question–statement distinction examined here.

6 Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Additionally, we extend our gratitude Jorg Dreyer for his support in running the experiment, and to our participants for their participation. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, project "Audio-visual prosody of (semi-)whispered speech" ZY 117/4-1).

References

M. W. Alibali, D. C. Heath, and H. J. Myers. 2001. Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44(2):169–188.

T. Baltrušaitis, C. Ahuja, and L. P. Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

A. Bangerter. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419.

J. Barker and S. Kita. 2026. The addressee's face as a cue for gesture production. *Language, Cognition and Neuroscience*, pages 1–18.

D. Bates, R. Kliegl, S. Vasishth, and H. Baayen. 2015. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

D. Bates, M. Maechler, B. Bolker, and S. Walker. 2020. Package 'lme4' (version 1.1-26) [computer software]. Comprehensive R Archive Network.

J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520.

J. B. Bavelas, N. Chovil, D. A. Lawrie, and A. Wade. 1992. Interactive gestures. *Discourse Processes*, 15(4):469–489.

P. Boersma and D. Weenink. 2019. Praat: Doing phonetics by computer, version 6.0.55. https://www.praat.org.

C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. 1996. About the relationship between eyebrow movements and f0 variations. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 4, pages 2175–2178.

M. Cruz, M. Swerts, and S. Frota. 2015. Variation in tone and gesture within language. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, pages 1–5.

E. Cvejic, J. Kim, C. Davis, and G. Gibert. 2010. Prosody for the eyes: Quantifying visual prosody using guided principal component analysis. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1433–1436.

J. P. De Ruiter. 1998. *Gesture and speech production*. Ph.D. thesis, Radboud University Nijmegen, Nijmegen.

J. P. De Ruiter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, 8(2):124–127.

J. P. De Ruiter, A. Bangerter, and P. Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.

M. Dohen and H. Loevenbruck. 2008. Audiovisual perception of prosodic contrastive focus in whispered french. *Journal of the Acoustical Society of America*, 123(5):3460.

K. Emmorey and S. Casey. 2001. Gesture, thought and spatial language. *Gesture*, 1(1):35–50.

S. Goldin-Meadow. 2009. How gesture promotes learning throughout childhood. *Child Development Perspectives*, 3(2):106–111.

M. Grice, S. Baumann, and R. Benzmüller. 2005. German intonation in autosegmental-metrical phonology. In S.-A. Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, Oxford, UK.

I. Guaïtella, S. Santi, B. Lagrue, and C. Cavé. 2009. Are eyebrow movements linked to voice variations and turn-taking in dialogue? an experimental investigation. *Language and Speech*, 52(2-3):207–222.

U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose. 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.

D. House. 2002. Intonational and visual cues in the perception of interrogative mode in swedish. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP/Interspeech)*, pages 1957–1960.

C. T. Ishi, H. Ishiguro, and N. Hagita. 2007. Analysis of head motions and speech in spoken dialogue. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 670–673.

R. M. Krauss, R. A. Dushay, Y. Chen, and F. Rauscher. 1995. The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology*, 31(6):533–552.

R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve. 2018. Package 'emmeans' (version 4.0-3) [computer software]. Comprehensive R Archive Network. http://cran.r-project.org/package=emmeans.

N. Mahdinazhad Sardhaei, M. Żygis, and H. Sharifzadeh. 2022. How do our eyebrows respond to masks and whispering? the case of persians. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2023–2027.

N. Mahdinazhad Sardhaei, M. Żygis, and H. Sharifzadeh. 2024. Facial expressions in different communication settings: A case of whispering and speaking with a face mask in farsi. *Language and Cognition*, 16(4):1639–1673.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

A. Melinger and W. J. Levelt. 2004. Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141.

L. Mol, E. Krahmer, A. Maes, and M. Swerts. 2011. Seeing and being seen: The effects on gesture production. *Journal of Computer-Mediated Communication*, 17(1):77–100.

K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137.

B. Rimé. 1982. The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, 12(2):113–129.

W. C. So, S. Kita, and S. Goldin-Meadow. 2009. Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33(1):115–125.

M. Swerts and E. J. Krahmer. 2004. Congruent and incongruent audiovisual cues to prominence. In *Proceedings of the 2nd International Conference on Speech Prosody*, pages 69–72.

F. Tao and C. Busso. 2014. Lipreading approach for isolated digits recognition under whisper and neutral speech. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1154–1158.

I. Van der Sluis. 2005. *Multimodal reference*. Ph.D. thesis, Katholieke Universiteit Brabant, Tilburg. Unpublished Doctoral Dissertation.

I. Van der Sluis and E. Krahmer. 2007. Generating multimodal references. *Discourse Processes*, 44(3):145–174.

I. F. Van der Sluis and E. J. Krahmer. 2004. The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, volume 4.

B. Winter. 2020. *Statistics for Linguists: An Introduction Using R*. Routledge.

H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. 2002. Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics*, 30(3):555–568.

M. Żygis and S. Fuchs. 2023. Communicative constraints affect oro-facial gestures and acoustics: Whispered vs normal speech. *Journal of the Acoustical Society of America*, 153(1):613–626.

Exploring Multimodal Turn-Taking Cues in Face-to-Face Conversation using Voice Activity Projection

Willem Berner
Lund University
willem.berner@gmail.com

Julio Cesar Cavalcanti
KTH Royal Institute of Technology
jcca@kth.se

Kalle Åström
Lund University
karl.astrom@math.lth.se

Gabriel Skantze
KTH Royal Institute of Technology
skantze@kth.se

Abstract

Turn-taking is a fundamental component of spoken interaction, and while humans naturally rely on both verbal and non-verbal signals, dialogue systems usually depend on audio cues alone. This paper investigates whether visual features from face-to-face conversations can enhance turn-taking prediction beyond what is achievable from audio-only. We extend the Voice Activity Projection (VAP) model, a self-supervised transformer-based model for predicting future voice activity, by incorporating visual features extracted from the large-scale Meta Seamless Interaction dataset of dyadic face-to-face conversations. The visual features include gaze direction, head movement, body and hand pose, and facial action units (FAU). For incorporating the visual features, we explore concatenation, cross-attention fusion, delta features, and trainable gating mechanisms. Results show that visual information improves performance over the audio-only baseline, with FAU being significantly more informative than other feature groups. Body and gaze features nevertheless contribute complementary information, as the model combining all features performs best. Furthermore, results indicate that performance on specific tasks varies depending on whether training and test data come from improvised (acted) or naturalistic (non-acted) conversations.

1 Introduction

Turn-taking in conversation appears effortless between humans, yet relies on a rich interplay of verbal and non-verbal cues across both visual and auditory modalities (Cowley, 1998; Skantze, 2021; Duncan, 1972). Turn-taking behaviour also varies substantially across individuals and dyads, suggesting that conversational timing is also affected by interaction-level dynamics (Cavalcanti and Skantze, 2025). While telephone conversations rely solely on auditory cues, such as lexical content and prosody, face-to-face interactions additionally incorporate visual cues, including gaze direction and gestures, to coordinate turn transitions (Duncan, 1972).

In spoken dialogue systems, turn-taking has traditionally been managed with a silence threshold. While simple, this approach often leads to interruptions or sluggish responses. Transformer-based models have significantly improved prediction of turn-taking, in particular the Voice Activity Projection (VAP) model (Ekstedt and Skantze, 2022b), which uses self-supervised learning on raw audio to predict which speaker(s) will be active over the next two seconds. VAP has been shown to outperform silence-threshold-based models for turn-taking (Skantze and Irfan, 2025).

A natural question is whether visual information can further improve such an audio-only model. Prior work has explored this, but on small or video-conferencing datasets (Saga and Pelachaud, 2025; O'Connor Russell and Harte, 2025a). Video-conferencing settings differ substantially from face-to-face interaction: turn transitions are shorter and more frequent in person, and gaze estimation is inherently ambiguous on-screen (Tian et al., 2023; Horstmann and Linke, 2022). Findings from such conditions may therefore not generalise to face-to-face dialogue

The release of the Meta Seamless Interaction dataset (Agrawal et al., 2025), shown in Figure 1, provides an opportunity to address this gap. It contains 4000 hours of recorded face-to-face conversations with paired audio, video, and pre-extracted visual features. In this work, we propose an architecture of VAP enhanced with visual features, trained on approximately 400 hours of data.

Concretely, our main question is: *Can visual cues from face-to-face conversations improve turn-taking prediction beyond audio alone, and, if so, which feature groups or individual features contribute most?* As a secondary exploratory analy-

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 50–58
October 29, 2026 ©2026 Association for Computational Linguistics



Figure 1: Two participants from the Seamless Interaction dataset.

sis, we examine predictive performance within and across the corpus's *improvised* and *naturalistic* subsets. These comprise prompted role-play involving actors and prompted, unscripted interactions between non-actors, respectively. We therefore ask: *How does turn-taking prediction performance vary when models are trained and evaluated within and across the improvised and naturalistic subsets?*

2 Related Work

Turn-taking is governed by a systematic organisation of cues that signal when a speaker intends to yield, hold, or take a turn (Sacks et al., 1974; Duncan, 1972). While prosody and timing are strong predictors (Cowley, 1998; Ekstedt and Skantze, 2022a), face-to-face interlocutors also rely heavily on visual signals such as gaze and gesture (Duncan, 1972; Kendon, 1967; ter Bekke et al., 2024). Turn-taking dynamics further differ across communication settings: transitions are shorter and overlaps more frequent in person than in mediated, on-screen conversation (Tian et al., 2023), motivating the study of visual cues specifically in large-scale face-to-face data.

Ekstedt and Skantze (2022b) introduced VAP, a self-supervised framework in which a transformer is trained to predict future binary voice-activity bins for two speakers, operating on raw audio via a frozen CPC encoder (van den Oord et al., 2018). VAP has since been widely adopted for turn-taking prediction, extended to real-time and continuous operation (Inoue et al., 2024b), as well as multilingual settings (Inoue et al., 2024a), and forms the basis of our work.

Several studies have extended VAP with visual information. Onishi et al. (2023) were the first to do so, augmenting VAP with non-verbal features – gaze, facial action units (FAU), head pose, and upper-body joints – on the NoXi corpus of mediated expert-novice interactions. They reported that FAU contributed most to turn-taking prediction. Saga and Pelachaud (2025) created a similar model that was trained on the French subset of NoXi (7 hours). They replaced hand-crafted action-unit inputs with a pre-trained facial-expression encoder; their models were competitive with, and on shift- and backchannel-prediction better than, prior multimodal baselines, though the small dataset limits its generalisation. O'Connor Russell and Harte (2025a) introduced MM-VAP, combining speech with facial expression, head pose, and gaze, and evaluated it on a 710-hour subset of the CANDOR video-conferencing corpus, again finding facial-expression features the most informative and outperforming an audio-only baseline on hold/shift prediction.

Across these studies, the FAU are consistently the most predictive, but all prior multimodal VAP work is confined to either small corpora (Onishi et al., 2023; Saga and Pelachaud, 2025) or video-conferencing recordings (O'Connor Russell and Harte, 2025a). Whether these findings hold for large-scale, face-to-face in-person (co-situated) conversation remains open. We address this using the Meta Seamless Interaction dataset (Agrawal et al., 2025), evaluating which visual feature groups improve turn-taking prediction at scale.

Beyond the VAP framework, other work has pursued multimodal turn-taking prediction with independently designed architectures. Wang et al. (2024) combined a neural acoustic model with a large language model, fusing lexical and acoustic representations to predict turn-taking. The experiments demonstrated that this multi-modal fusion of lexical and acoustic information consistently outperforms baseline models relying on a single modality. Their approach, however, is limited to audio and text and does not incorporate visual signals. Lin et al. (2025) extended this line of work to a third modality, proposing a tri-modal architecture with text, audio and video, fusing separate pretrained backbones (GPT-2, HuBERT, VideoMAE) through a flexible fusion module. Their model frames the task as discrete classification of keep, turn-taking, and backchannel actions. To support this work, the authors also introduced MM-F2F, a tri-modal (text, audio, video) dataset with both turn-taking

and backchannel annotations, comprising over 210 hours of face-to-face conversation.

3 Data

3.1 Seamless Interaction Dataset

We use the open-source Meta Seamless Interaction dataset (Agrawal et al., 2025), which comprises over 4000 hours of face-to-face conversations between two participants. Recordings include two separated mono audio channels and corresponding video files, with pre-extracted visual features provided per frame. The corpus distinguishes improvised role-play involving actors from naturalistic prompted interactions involving non-actors. The two variants are labelled *improvised* and *naturalistic* in the dataset. We acknowledge that naturalistic interaction is not to be treated as equivalent to unelicited everyday conversation. We use a subset of approximately 400 hours total, split into training (328h), validation (41h), and test (41h) sets with a balanced improvised/naturalistic distribution in each split (Table 1). The three sets are conversation-disjoint but not participant- nor dyad-disjoint: dyads in the test set also occur in training and validation.

Split	Total	Improvised	Naturalistic
Train	328h	47.3%	52.7%
Val	41h	46.3%	53.7%
Test	41h	46.3%	53.7%

Table 1: Dataset partitions of our dataset.

3.2 Visual Features

We use four groups of visual features:

Gaze features ($\mathbb{R}^2$ per frame): pitch and yaw of gaze direction per speaker, extracted by a Meta-internal tool. Values are noisy when eyes are partially occluded.

Head features ($\mathbb{R}^3$ per frame): pitch, yaw, and roll of head orientation.

Body and hand features: Body pose is represented as 21 joints ($b_i \in \mathbb{R}^3$) from the SMPL-H model (Romero et al., 2017), estimated via HMR2.0 (Goel et al., 2023). Hand pose is represented as 15 joints ($h_i \in \mathbb{R}^3$) per hand, estimated with ViTPose (Xu et al., 2022).

Facial Action Units (FAU): 24 action units from the Facial Action Coding System (FACS) (Ekman and Friesen, 1978), each with intensity values $\in [0, 1]$. Gaze-related units are excluded as gaze is already represented separately.

We define the following feature subsets for ablation: *All features* (body, hands, gaze, FAU); *Body features* (body, hands); *Gaze features* (gaze, head); *FAU features* (FAU only). The full visual feature vector has dimension 182.

The visual features are supplied at 30 Hz; we linearly interpolate them to match the baseline's 50 Hz rate.

4 Methodology

4.1 Baseline: Voice Activity Projection

The VAP model (Ekstedt and Skantze, 2022b) encodes raw audio via a frozen CPC encoder (van den Oord et al., 2018), producing 256-dimensional latent representations at 50 Hz. For each of the two speakers, a self-attention transformer models within-speaker context, producing representations X_a and X_b for speakers A and B, respectively. This is followed by a bidirectional cross-attention between speakers

$$\hat{X}_a = \text{CrossAttn}(X_a, X_b),$$
$$\hat{X}_b = \text{CrossAttn}(X_b, X_a).$$

All attention operations use four heads. A combinator merges the two enriched representations:

$$y = \text{GELU}(\text{LayerNorm}(W_a \hat{X}_a + W_b \hat{X}_b)).$$

The VAP head predicts the probability distribution over all 256 combinations of 8 binary future voice activity bins (4 per speaker, spanning 2 seconds). Bins are spaced as $[(0.0, 0.2); (0.2, 0.6); (0.6, 1.2); (1.2, 2.0)]$ seconds, giving finer resolution for the near future. For a bin to be considered as voice active, the majority of it must be covered (Figure 2).

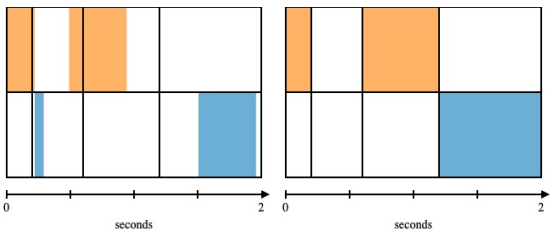


Figure 2: Voice activity before (left) and after binning (right). Each bin is active if the majority of its duration contains speech. Upper row: speaker A; lower row: speaker B.

Model	Loss	H/S BA (%)	Sh/L BA (%)	Ph*/Ps BA (%)
M0: Baseline	2.166	76.41	84.11	63.90
M1: Concatenation	2.122	77.87	84.67	63.30
M2: Cross-attention	2.119	77.27	84.80	64.30
M3: Cross-attention + Δ_3	2.107	77.83	84.81	64.51
M4: Cross-attention + Δ_{10}	2.130	77.76	84.20	65.38
M5: Cross-attention + Δ_3 + Early Gates	**2.100**	78.31	**85.26**	65.14
M6: Cross-attention + Δ_3 + Late Gates	2.106	**78.81**	84.67	**65.51**

Note. To assess statistical significance, we apply paired bootstrapping (1000 samples of 694 test conversations with replacement): **(M0 vs M5) 95% CI**: Loss: [-0.0691, -0.0616], H/S: [1.22, 2.60], S/L [0.59, 1.74], Ph/Ps [0.83, 2.03], i.e., all differences are significant. The asterisk refers to the subsample of Ph described in Section 4.3.

Table 2: Balanced accuracy (BA) and VAP loss for all models on the full test set. Bold indicates best performance per metric.

The voice activity (VA) head predicts current VA via a sigmoid. The combined training loss is $\mathcal{L} = \mathcal{L}_{\text{VAP}} + \mathcal{L}_{\text{VA}}$, with model selection based on $\mathcal{L}_{\text{VAP}}$ alone. The full architecture of the VAP model is shown in Figure 4.

4.2 Multimodal Extensions

We extend the baseline through three sequential design decisions: how to fuse the two modalities (M1, M2), how to add motion through delta features (M3, M4), and where to apply trainable gates (M5, M6). Each decision is a choice between two variants; we keep the lower-loss variant and build the next decision on top of it. Table 2 shows the results of the configurations.

Concatenation/Cross-attention In M1 (concatenation), the visual feature vector (dimension 182) is concatenated with the audio representation (dimension 256) per speaker, then projected back to 256 dimensions via a linear layer:

$$\tilde{x} = W_p[x; v] + b_p, \quad W_p \in \mathbb{R}^{256\times 438}.$$

This output is then passed as an input to the audio-only architecture.

In M2 (cross-attention), visual features are first linearly projected to dim 256, processed by self-attention, and then fused with audio via a two-step cross-attention. First, visual information is conditioned on audio:

$$\tilde{X}_v = \text{CrossAttn}(X_v, X_a),$$

then audio is updated using the audio-conditioned visual information:

$$\tilde{X}_a = \text{CrossAttn}(X_a, \tilde{X}_v).$$

The intuition is that audio-conditioned visual features provide a more informative key/value source for refining the audio representation.

Delta3/Delta10 Since the visual features encode static position rather than motion, we augment them with delta features:

$$\Delta(\text{feat}_i)_j = (\text{feat}_i)_j - \frac{1}{N}\sum_{k=j-N}^{j-1} \text{feat}_k,$$

capturing deviation from a short-term mean. We test $N = 3$ (M3) and $N = 10$ (M4). Adding delta features doubles the vector to dimension 364, which we map back to 256.

Early/Late gates A gate vector $g = \sigma(W_g x + b_g)$, with weights initialised to zero ($g = 0.5\ \forall x$ initially), scales features element-wise: $\bar{x} = g \odot x$. We try placing gates before the visual self-attention (early gates, M5) or before cross-attention (late gates, M6). The M5 model is shown in Figure 4.

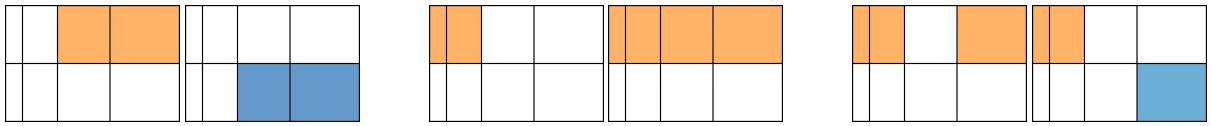
4.3 Evaluation Metrics

Beyond VAP loss, we evaluate on three interpretable turn-taking tasks (Ekstedt and Skantze, 2022b). Examples of the events are shown in Figure 3.

Hold/Shift (H/S): accuracy at predicting whether a mutual silence ($\geq$100 ms, with 1 s pre/post single-speaker constraint) is followed by the current speaker continuing (Hold) or the other speaker taking over (Shift). This is perhaps the most central task for conversational systems to know whether the user is yielding the turn or not.

Short/Long (Sh/L): accuracy at predicting from the first 200 ms of a new speaker onset whether the utterance will be short ($<$1.5 s) or long ($\geq$1.5 s). For conversational systems, this can be useful to predict whether the user has just started a backchannel, in which case the system might continue, or is attempting to initiate a longer turn.

Shift Prediction (Ph/Ps): accuracy at predicting an upcoming turn shift while the current speaker is



(a) Hold vs. Shift (b) Short vs. Long (c) Prehold vs. Preshift

Figure 3: The three event types illustrated: hold vs. shift (given that person 1, orange, had been speaking prior to the window shown, whether they keep the floor or yield it), short vs. long (utterance duration), and prehold vs. preshift (activity preceding a hold/shift). Example configurations; other valid ones exist.

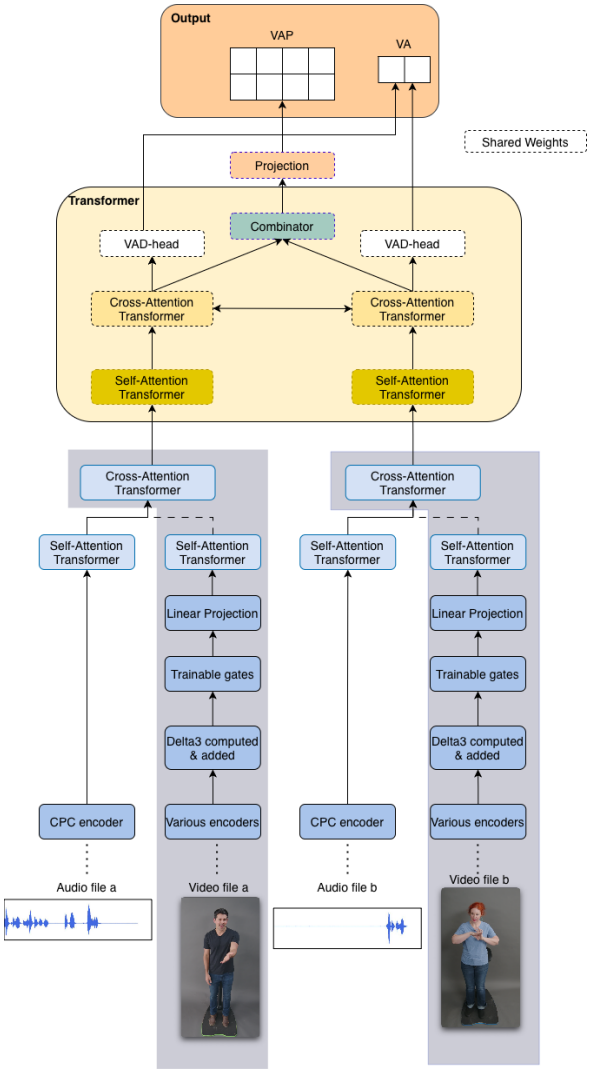


Figure 4: M5 architecture. Shaded blocks show components added to the M0 audio-only backbone, where two CPC-encoded streams pass per-channel self-attention, cross-speaker attention, and a combinator to predict VAP and VA. M5 adds visual features, augmented with deltas (Δ_3), passed through early gates and self-attention, then cross-attended to audio before the main transformer.

still active, measured as pre-shift (Ps) versus pre-hold (Ph). Due to heavy class imbalance, pre-hold events are subsampled to match pre-shift counts; we report the average over $N = 10$ subsampling seeds.

In addition to subsampling pre-holds, all tasks are evaluated using balanced accuracy (BA) to account for smaller class imbalances.

Training used the Adam optimizer on an NVIDIA RTX 3090, the same learning rate as the baseline VAP, and early stopping with a patience of 5 epochs.

5 Results

5.1 Model Comparison

Table 2 summarises results for the proposed architectures on the full test set, all using all features.

While the differences are not very big, every multimodal model reduces VAP loss compared to the baseline (M0: 2.166). M5 achieves the lowest loss (2.100) and is selected as the best architecture.

Cross-attention (M2) outperforms concatenation (M1) on loss and most metrics, though the difference is modest. Adding Δ_3 features (M3) reduces loss further, indicating that motion information complements positional features. Gates (M5/M6) give further gains, as adaptive weighting helps the model suppress uninformative dimensions.

5.2 Feature Group Ablation

Using M5 architecture, we train separate models for each feature subset (Table 3). All Features achieve the best results overall, while FAU achieve the best result for the single feature groups.

Features	Loss	H/S	Sh/L	Ph*/Ps
Baseline	2.166	76.41	84.11	63.90
All	2.100	78.31	85.26	65.14
Body	2.143	75.51	83.97	63.64
Gaze	2.150	76.18	84.03	62.49
FAU	**2.115**	**77.52**	**84.89**	**63.95**

Note. The asterisk refers to the subsample of Ph.

Table 3: Feature group ablation using M5 architecture. Loss in absolute numbers balanced accuracy in %.

Contrast	Metric	Mean Δ	95% CI
FAU vs body	Loss	−0.028	**[-0.032, -0.025]**
	H/S	2.01	**[1.27, 2.66]**
	Sh/L	0.92	**[0.34, 1.46]**
	Ph/Ps	0.55	[−0.18, 1.23]
FAU vs gaze	Loss	−0.035	**[-0.038, -0.032]**
	H/S	1.34	**[0.65, 2.05]**
	Sh/L	0.85	**[0.31, 1.33]**
	Ph/Ps	1.56	**[0.94, 2.24]**
Body+FAU vs FAU	Loss	−0.009	**[-0.010, -0.007]**
	H/S	0.74	**[0.22, 1.29]**
	Sh/L	0.17	[−0.23, 0.57]
	Ph/Ps	1.00	**[0.43, 1.56]**
Gaze+FAU vs FAU	Loss	−0.004	**[-0.006, -0.001]**
	H/S	0.63	**[0.04, 1.23]**
	Sh/L	0.18	[−0.28, 0.63]
	Ph/Ps	−0.38	[−0.97, 0.51]

Table 4: Paired bootstrap differences for feature-complementarity contrasts. Balanced accuracy differences in percentage points while loss differences are absolute. For balanced accuracy, positive values indicate better performance by the first model in the contrast. For loss, negative values indicate lower loss for the first model. Confidence intervals that exclude zero are highlighted in bold.

In Table 4, the differences between feature groups are assessed more closely. For FAU vs Body, three out of the four confidence intervals exclude zero; the only non-significant difference is for Ph/Ps. All four confidence intervals for FAU vs Gaze are significant. Both Body+FAU and Gaze+FAU outperform FAU alone on some metrics: Body+FAU has three CIs excluding zero, Gaze+FAU has two. This points to Body and Gaze providing complementary information when combined with FAU, even if their independent contribution is limited. It appears that body is particularly informative for H/S and Ph/Ps, and gaze for H/S.

5.2.1 Individual FAU Importance

We focus the finer-grained analysis on FAU because the feature-group comparison identified them as the strongest visual group overall (Tables 3-4). To estimate the contribution of individual action units, we zero out each FAU (and its delta) at inference and measure the resulting loss increase (Table 5).

JawDrop (AU26), BrowLowerer (AU4), and EyesClosed (AU43) are the most important individual units. These results are intuitively plausible: jaw movement is directly tied to speaking, and eye closure relates to timing and emphasis. However, we acknowledge that the FAU are likely highly correlated, and that this approach only shows influence when one is zeroed at a time.

Action Unit	Loss	Relative increase
None (full model)	2.115	–
AU26 (JawDrop)	2.130	+0.71%
AU4 (BrowLowerer)	2.129	+0.66%
AU43 (EyesClosed)	2.126	+0.52%
AU12 (LipCornerPull)	2.124	+0.43%
AU10 (UpperLipRaiser)	2.122	+0.33%
(remaining AUs: $\leq 0.14\%$ increase)		

Table 5: Loss increase when zeroing out single FAU. Top-5 most impactful shown.

Table 6 shows the value for each of the most influential FAU for the frame right before a hold/shift event, together with its 95% confidence interval, where P1 is the active speaker. It indicates that P1 is more likely to close their eyes (AU43) and have their mouth open (AU26) when holding the turn, rather than yielding it. Also, P1 is more likely to smile (AU12) when yielding the turn. P2 is more likely to open their mouth (AU10, AU26) and smile (AU12) when they are about to take the turn, while they tend to lower their eyebrows (AU4) when the other participant holds the turn.

5.3 Improvised vs. Naturalistic

The M5 model (all features) achieves a lower, i.e. better, VAP loss on naturalistic conversations than on improvised ones (1.968 vs. 2.203). The pattern reverses for the event-based metrics: improvised conversations yield higher balanced accuracy on hold/shift (78.86% vs. 77.89%) and short/long (85.71% vs. 83.90%), while the pre-hold/pre-shift difference is small. We assess significance using stratified bootstrapping, resampling improvised and naturalistic subsets separately and computing the improvised−naturalistic difference. Only the loss difference is significant ($\Delta = 0.235$, CI $[0.148, 0.324]$), confirming that the model predicts naturalistic dynamics more accurately overall. The balanced-accuracy differences — H/S ($\Delta = 0.88$, CI $[-1.08, 2.75]$), S/L ($\Delta = 1.73$, CI $[-0.42, 3.96]$), and Ph/Ps ($\Delta = 0.14$, CI $[-2.19, 2.55]$) — all have CIs spanning zero, so the trend toward higher event accuracies on improvised data is not significant. One explanation is that improvised conversations have more exaggerated turn-taking cues, making discrete events easier to classify, whereas naturalistic conversations have smoother dynamics that the model captures more

	P1		P2	
AU	Shift	Hold	Shift	Hold
AU26 JawDrop	**0.058 [0.054, 0.062]**	**0.066 [0.064, 0.067]**	**0.048 [0.045, 0.051]**	**0.036 [0.035, 0.038]**
AU4 BrowLowerer	0.367 [0.353, 0.380]	0.359 [0.352, 0.366]	**0.363 [0.349, 0.377]**	**0.416 [0.409, 0.422]**
AU43 EyesClosed	**0.349 [0.332, 0.367]**	**0.394 [0.385, 0.403]**	0.317 [0.301, 0.334]	0.332 [0.324, 0.340]
AU12 LipCornerPull	**0.525 [0.507, 0.543]**	**0.390 [0.383, 0.397]**	**0.511 [0.494, 0.529]**	**0.455 [0.448, 0.463]**
AU10 UpperLipRaiser	0.352 [0.338, 0.365]	0.350 [0.344, 0.356]	**0.363 [0.350, 0.376]**	**0.275 [0.270, 0.281]**

Table 6: Action unit averages, with confidence intervals, for the last frame before each event, for the active participant 1 (P1) and the passive participant 2 (P2) under Shift and Hold conditions.

Domain	Shift/min	Hold/min	Shift/Hold
Improvised	1.10	4.68	0.24
Naturalistic	1.41	6.56	0.21

Table 7: Shift and hold rates by domain per minute, and Shift/Hold ratio.

accurately at the level of overall voice-activity distribution. This remains an open question for future work.

Table 7 shows the shift and holds in the naturalistic and improvised test-sets. Naturalistic conversations show more shifts and holds per minute, indicating more frequent pauses, whereas speakers in improvised conversation produce longer inter-pausal units.

Table 8 shows how models perform on partitions of the test set, when trained on naturalistic or improvised. Models trained on improvised data appear to perform better on shift, while the models trained on naturalistic appear to perform better on hold. This might imply that the improvised models lean towards predicting shift while the naturalistic models tend to predict hold, disregarding test-set. The same directional bias appears on the other tasks. On pre-hold/pre-shift, improvised-trained models favour pre-shift while naturalistic-trained models favour pre-hold, which follows directly from the shift/hold ratio (Table 7). On short/long, improvised-trained models are prone to predict long, while naturalistic-trained models favour short; this is consistent with the same account, although the underlying short/long base rates are not reported here. Both effects are smaller than for hold/shift. We present these as descriptive trends and do not draw significance-based conclusions from them.

6 Discussion

Every multimodal variant lowers VAP loss relative to the audio-only baseline (Table 2), and for our best model the gains are statistically significant across loss and all three pre-defined tasks. This confirms that visual features add information that audio alone does not capture.

Concatenation (M1) and cross-attention (M2) perform almost identically on loss and on the event-based metrics. We anticipated that cross-attention, with its explicit conditioning between modalities, would show a clearer advantage. One interpretation is that the audio transformer integrates the visual features regardless of how they enter the network; another is that at this data scale the fusion mechanism matters less than the presence of the visual signal itself. We cannot separate these explanations from the present experiments.

Among the feature groups, FAU are the most informative. Alone, they come close to the performance of the all features model and clearly outperform body and gaze (Table 3), a gap that is significant for loss and for the hold/shift and short/long tasks (Table 4). The same ranking has been reported for mediated and video-conferencing data (Onishi et al., 2023; O'Connor Russell and Harte, 2025a); our results show that it also holds for large-scale face-to-face conversation. Body and gaze add little in isolation, yet the full model still improves over any single group, so these features carry information that is complementary rather than redundant.

The size of the gains is itself informative. On clean, channel-separated audio the acoustic signal already carries most of what the task requires, so visual features are better understood as supplementing audio than replacing it. They help most where the acoustic evidence is ambiguous, which fits the finding that facial action units are the strongest visual contributor: jaw, mouth, eye, and brow move-

			Hold/Shift			Short/Long			Pre-Hold/Pre-Shift		
Train	Test	Loss	BA	H	S	BA	Sh	L	BA	Ph*	Ps
Imp	Imp	2.235	79.41	**87.20**	**71.63**	84.45	**73.52**	**95.39**	66.46	**94.25**	**38.66**
Imp	Nat	2.044	77.75	**84.05**	**71.45**	80.61	**64.66**	**96.55**	66.80	**93.16**	**40.44**
Nat	Imp	2.304	77.80	**89.37**	**66.23**	84.50	**77.42**	**91.58**	64.63	**95.19**	**34.07**
Nat	Nat	2.009	77.04	**88.42**	**65.65**	82.15	**70.53**	**93.77**	64.98	**96.09**	**33.87**

Note. The asterisk refers to the subsample of Ph described in Section 4.3.

Table 8: Cross-testing results between improvised and naturalistic training and test sets. BA denotes balanced accuracy.

ment are tied directly to speech onset and offset and to communicative intent. This also implies that the gains reported here may be conservative: visual access to the face and lips raises speech intelligibility in noise (Sumby and Pollack, 1954), and a turn-taking model evaluated under added noise loses much of its audio-only hold/shift accuracy while a multimodal model degrades far less (O'Connor Russell and Harte, 2025b). The recordings used are clean and the two speakers occupy separate channels, so the contribution of visual features may be larger under noisier, single-channel conditions typical of deployed dialogue systems.

7 Conclusion

Visual features from large-scale face-to-face conversations improve turn-taking prediction over the audio-only VAP. Our best model (M5: cross-attention with Δ_3 features and early gates) shows statistically significant improvements across all three evaluation tasks. Among individual feature groups, FAU are the most informative, though combining all features performs best. Also, the results appear to vary if trained or tested on acted or non-acted conversations. These results support the use of visual information in spoken dialogue systems aiming for more natural turn-taking behaviour.

8 Limitations

Several limitations qualify these results. Gaze estimates are unreliable when the eyes are partly occluded, which likely caps the contribution of the gaze group. A single averaging window N for the delta features is unlikely to suit feature types with different temporal dynamics. The multimodal model is also more costly to run than the audio-only baseline, a consideration for online use (Inoue et al., 2024b). Promising directions include end-to-end video encoders that remove the need for hand-engineered feature extraction, and per-group tuning of the delta window N.

A final limitation concerns speaker and dyad identity. Turn-taking timing is highly dyad-specific: Cavalcanti and Skantze (2025) find that idiosyncratic and "dyadosyncratic" factors dominate variation in transition timing. This affects our setup in two ways. Our splits are conversation-disjoint but not dyad-disjoint, so the model may exploit dyad-specific regularities seen in training; the reported scores should thus be read as an upper bound, pending a dyad-disjoint evaluation. Because the model is trained pooled across dyads with no speaker or dyad conditioning, its accuracies are averages over a heterogeneous population. Conditioning on, or adapting to, the speaker or dyad is a natural direction for future work.

9 Acknowledgements

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Grants #2024/06797-2, and #22/05848-7.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeffrey M. Girard, Martin Gleize, and 64 others. 2025. Seamless Interaction: Dyadic Audiovisual Motion Modeling and Large-Scale Dataset. *Preprint*, arXiv:2506.22554.

Julio Cesar Cavalcanti and Gabriel Skantze. 2025. "Dyadosyncrasy", Idiosyncrasy and Demographic Factors in Turn-Taking. In *Proceedings of Interspeech 2025*, pages 1093–1097. ISCA.

Stephen J. Cowley. 1998. Of Timing, Turn-Taking, and Conversations. *Journal of Psycholinguistic Research*, 27(5):541–571.

Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, USA.

Erik Ekstedt and Gabriel Skantze. 2022a. How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice Activity Projection: Self-Supervised Learning of Turn-Taking Events. In *Proceedings of Interspeech 2022*, pages 5190–5194. ISCA.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794. IEEE.

Gernot Horstmann and Linda Linke. 2022. Perception of direct gaze in a video-conference setting: the effects of position and size. *Cognitive Research: Principles and Implications*, 7(67).

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual Turn-taking Prediction Using Voice Activity Projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11873–11883. European Language Resources Association (ELRA) and ICCL.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection. *Preprint*, arXiv:2401.04868.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Yuxin Lin, Yinglin Zheng, Ming Zeng, and Wangzheng Shi. 2025. Predicting Turn-Taking and Backchannel in Human-Machine Conversations Using Linguistic, Acoustic, and Visual Signals. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15310–15322. Association for Computational Linguistics.

Sam O'Connor Russell and Naomi Harte. 2025a. Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 209–221. Association for Computational Linguistics.

Sam O'Connor Russell and Naomi Harte. 2025b. Visual Cues Support Robust Turn-taking Prediction in Noise. In *Proceedings of Interspeech 2025*, pages 1073–1077. ISCA.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*, pages 13–21. Association for Computing Machinery.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Takeshi Saga and Catherine Pelachaud. 2025. Voice Activity Projection Model with Multimodal Encoders. *Preprint*, arXiv:2506.03980.

Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67(101178).

Gabriel Skantze and Bahar Irfan. 2025. Applying General Turn-taking Models to Conversational Human-Robot Interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 859–868. IEEE Press.

William H. Sumby and Irwin Pollack. 1954. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2):212–215.

Marlijn ter Bekke, Linda Drijvers, and Judith Holler. 2024. Hand Gestures Have Predictive Potential During Conversation: An Investigation of the Timing of Gestures in Relation to Speech. *Cognitive Science*, 48(1):e13407.

Ying Tian, Siyun Liu, and Jianying Wang. 2023. A Corpus Study on the Difference of Turn-Taking in Online Audio, Online Video, and Face-to-Face Conversation. *Language and Speech*, 67(3):593–616.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *Preprint*, arXiv:1807.03748.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-Taking and Backchannel Prediction with Acoustic and Large Language Model Fusion. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125. IEEE.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*, volume 35, pages 38571–38584. Neural Information Processing Systems Foundation, Inc.

Automatic Annotation of Co-speech Gesture for Multimodal Communication Research – A Framework and a Case Study

Ilya Burenko[1,2], Peter Uhrig[3,1], Anna Wilson[4]

[1]ScaDS.AI Dresden/Leipzig, Germany*
[2]Technische Universität Dresden, Dresden, Germany
[3]Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[4]Oxford School of Global and Area Studies, University of Oxford, United Kingdom
Correspondence: ilia.burenko@tu-dresden.de

Abstract

Large-scale multimodal communication research increasingly depends on computational methods capable of extracting reliable gesture annotations from video data. Yet it remains unclear to what extent automatically generated annotations can support the linguistic investigation of naturally occurring co-speech gesture and where expert interpretation remains indispensable. This paper presents a framework for the automatic analysis and annotation of co-speech gesture in media corpora, developed to support multimodal communication research at scale. We build our framework on top of the existing computer vision frameworks for keypoint extraction and active speaker detection, with post-processing steps carefully designed to mitigate noisy output. Building on the processed keypoint trajectories, we implement a configurable rule-based annotation system developed through close collaboration between computer vision researchers and linguists. We demonstrate the usefulness of the framework through a case study investigating vertical hand movement accompanying English expressions of future and past, examining which aspects of this gestural pattern can be captured through automatic annotation and which require expert multimodal interpretation. We find that at corpus scale, greater mean amplitude and higher mean peak velocity converge in *will*-future, suggesting greater kinematic prominence in the multimodal expression of futurity. More broadly, the case study demonstrates how computational pipelines can support, rather than replace, human analysis by enabling multimodal hypotheses to be investigated on datasets that would be impractical to analyse manually.

1 Introduction

The combination of increasingly large multimodal corpora and advances in computational methods is transforming multimodal communication research. Researchers can now investigate patterns of speech and gesture across thousands of naturally occurring interactions rather than a few dozen manually annotated examples. At the same time, this opportunity presents a major methodological challenge: producing reliable multimodal annotations at such a scale remains prohibitively time-consuming when performed manually, especially when the annotation process requires trained experts.

In linguistics, ELAN (Wittenburg et al., 2006) serves as the standard environment for producing and editing manual annotations of audiovisual recordings. It enables researchers to create multiple annotation tiers, which are stored alongside the media files. ELAN annotations can also be manipulated programmatically using the Python package pympi-ling (Lubbers and Torreira, 2013-2026). Larger collections of annotated data can be organised into corpora and queried using corpus managers such as CQPweb (Hardie, 2012).

Although the machine learning methods have achieved substantial improvements in computational performance, comparatively less attention has been paid to how they support multimodal communication research itself. Most existing work evaluates computational performance, whereas comparatively few studies investigate what kinds of linguistic hypotheses automatic annotation can reliably support, where expert multimodal interpretation remains indispensable, and what new research questions become accessible once annotation can be performed computationally at corpus scale.

To demonstrate the usefulness of the framework, we apply it to the NewsScape 2016 corpus (Uhrig, 2022), which contains more than 30,000 hours of American TV news (in a broad sense) with subtitles comprising over 230 million words that have been time-aligned with Gentle (Ochshorn and Hawkins, 2015). Processing a dataset of this size presents substantial computational challenges. Temporal

*Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 59–71
October 29, 2026 ©2026 Association for Computational Linguistics

expressions cannot simply be retrieved from raw video but require several computationally demanding processing stages, including scene detection, pose estimation, active speaker detection and keypoint extraction.

Rather than evaluating the framework solely as a technical system, we demonstrate its usefulness through a genuine multimodal communication research question: whether previously identified patterns of vertical gesture accompanying English temporal conceptualisation generalise to a substantially larger automatically annotated corpus. The case study simultaneously investigates what aspects of this phenomenon current automatic annotation can recover reliably, where expert multimodal interpretation remains essential, and what additional linguistic questions become accessible once gesture can be analysed computationally at corpus scale.

The paper therefore makes three contributions. First, it presents a scalable, modular framework for the automatic annotation of co-speech gesture in multimodal corpora. Secondly, through a multimodal case study, it demonstrates how computational annotation can generalise findings previously established through manual multimodal analysis while revealing both the capabilities and the current limitations of automatic gesture annotation. Thirdly, it shows how corpus-scale computational analysis enables new forms of multimodal communication research by quantifying kinematic properties of gesture that would be impractical to investigate manually.

The remainder of the paper is organised as follows. Section 2 reviews related work. Sections 3 and 4 describe active-speaker keypoint extraction, post-processing and the automatic annotation framework. Section 5 validates the automatic annotation against expert human annotation, and Section 6 presents the case study of co-speech gesture accompanying English expressions of future and past. Section 7 discusses the methodological and linguistic implications, followed by limitations of the automatic annotation and analysis in Section 8.

2 Related work

Scene Detection. Several stages of our pipeline require partitioning a video into scenes – segments of footage in which the camera position and filmed content remain relatively stable –, including active speaker detection (Section 3) and keypoint normalisation (Section 4). We use PySceneDetect[1] for this purpose, a content-aware detection tool that identifies a scene boundary whenever the rate of change in frame colour and saturation exceeds a fixed threshold. PySceneDetect exposes a range of configurable parameters, but we found the default settings sufficient for our dataset.

Pose Estimation. To detect people in video frames and estimate two-dimensional coordinates of their joints, we need to use a pose estimation method. There is a wide range of readily available pose estimation models, which allows to choose the best trade-off between speed and quality. Light-weight models such as MediaPipe (Lugaresi et al., 2019) can process video with high frame-per-second ratio, sacrificing some quality. In contrast, recent large models such as Sapiens (Khirodkar et al., 2024) yield better performance, but it comes at an additional computational cost. Among others we compared AlphaPose (Fang et al., 2022), HRNet (Cheng et al., 2020), or OpenPifPaf (Kreiss et al., 2021), and OpenPose (Cao et al., 2019). We chose OpenPose as the optimal tool to process the full 30,000-hour dataset as it can process multiple people on screen, is very robust, and proved to be the best trade-off between speed and quality. We refer the reader to Gao et al. (2025) for a recent review of pose estimation methods.

Active Speaker Detection. To extract OpenPose coordinates for the active speaker, we first automatically identify which individual is speaking at any given time. The AVA Challenge 2022[2] evaluated a number of approaches for active speaker detection. These employ diverse architectural strategies: some approaches, such as that proposed by (Köpüklü et al., 2021), utilize 3D convolutional networks combined with bidirectional Gated Recurrent Units (Cho et al., 2014) to capture temporal dependencies in the data, while alternative methods leverage graph networks to model spatio-temporal relationships between detected faces across sequential frames (Min et al., 2022). For a more detailed discussion, we refer the reader to the survey by Nur Aisyah Mohd Robi et al. (2024).

For our pipeline, we adopt TalkNet-ASD, which employs a transformer architecture with cross-attention mechanisms (Tao et al., 2021) to achieve robust active speaker detection; the authors' pub-

[1] https://github.com/Breakthrough/PySceneDetect

[2] https://research.google.com/ava/challenge.html

licly available implementation[3] is straightforward to run, test, and adapt. TalkNet-ASD partitions the input video into scenes and detects a face bounding box for each visible individual in every frame. Bounding boxes with substantial overlap across consecutive frames are merged into what the authors term a *track*. If the pipeline fails to detect a face for a given frame within a track, it linearly interpolates the bounding box coordinates for up to 10 frames by default.

Annotation Pipelines. To the best of our knowledge, there is no comparable framework for the automatic annotation of co-speech gestures. Envisionbox[4] collects community code snippets and recipes for the application of various computer vision tools and how to exploit their results, but no fully-integrated automatic pipeline is provided.

3 Adapting TalkNet-ASD to Dataset Scale

We adapt TalkNet-ASD (Tao et al., 2021) to accommodate the scale and heterogeneity of large multimodal datasets. Two properties of such datasets are not well supported by the original implementation. First, some videos contain exceptionally long scenes, sometimes extending to several hours; press conference recordings filmed from a fixed camera angle, for instance, produce very long tracking sequences. Second, video collections often span a broad range of resolutions, from low-resolution recordings dating back to the early 2000s to recent 4K content. The original implementation loads all video frames into memory, which is not feasible for high-resolution video.[5]

To address these limitations, we implement three optimisations. High-resolution videos are resized before being loaded into memory, reducing computational overhead. Long scenes and tracks are split into smaller sub-scenes and sub-tracks, which can be processed more efficiently. Finally, unlike the original implementation, which processes a single image per inference step, our version uses a batch size of 32, which accelerates inference considerably. These changes do not alter the underlying algorithm; they enable its use with video of any resolution and scenes of any length.

[3] https://github.com/TaoRuijie/TalkNet-ASD

[4] https://envisionbox.org/index.html

[5] Note, however, that the pre-trained network only accepts video at 25 fps. TalkNet-ASD converts any input video to 25 fps before applying the network.

4 ASD Keypoints Extraction

For body pose annotation, we use OpenPose (Cao et al., 2019), from which we extract keypoints in the BODY 25 scheme together with hand and face keypoints. The pipeline is not tied to OpenPose specifically: keypoints produced by any other pose estimation model can be used in its place, provided they are mapped onto the BODY 25 scheme, which is generally straightforward.

Since OpenPose estimates keypoints for every person visible in a frame, we need a way to pick out the active speaker among them. To identify which of these predictions corresponds to the active speaker, we use the bounding box predicted by TalkNet-ASD: for each frame, we select the individual whose head position falls within this bounding box.

The resulting keypoint time series are noisy in several ways. Keypoints are sometimes missing entirely, for example when motion blur or occlusion prevents OpenPose from detecting a body part. The coordinates also jitter slightly from frame to frame, even when the corresponding body part is not actually moving. Finally, because OpenPose predicts coordinates in raw pixel values, the same body part can appear at very different scales depending on the camera angle and the speaker's distance from the camera. Left uncorrected, these issues make the data difficult to work with, both for automatic analysis and for manual inspection. We therefore apply a three-step preprocessing pipeline, separately for each track (Section 2): interpolation and extrapolation, smoothing, and normalisation.

Missing values are handled through interpolation and extrapolation. Gaps between two valid measurements are filled by linear interpolation, while missing values at the start or end of a track are filled by extrapolating the nearest valid value. Further details are given in the Appendix C.

Jitter is reduced through smoothing. Each coordinate is replaced with a weighted average over a sliding window of 11 frames, centred on the current frame, using Epanechnikov kernel weights. This window size offered the best trade-off in our data between removing jitter and preserving genuine small movements.

To correct for scale, we normalise the coordinates into a shared coordinate space that is independent of the speaker's size and position in the frame. Body keypoints are centred on the nose and scaled by the track-averaged distance between the nose

and neck; face keypoints are centred on the tip of the nose and scaled by the track-averaged distance between the nasal bridge and the nose tip. Hand keypoints are left unnormalised, since OpenPose's hand predictions are comparatively unreliable to begin with.

Figure 1 shows the pipeline's effect: the top panel shows raw pixel coordinates, and the bottom panel shows the same time series after interpolation, extrapolation, smoothing, and normalisation. Implementation details, including the keypoint indices used and the reasoning behind our normalisation choices, are provided in Appendix C.

5 Annotation Validation

To validate the pipeline for our case study, which focuses on the vertical axis, we compared the automatic annotation with human annotations for the same data (24 episodes of SophieCo. Visionaries show; see Wilson et al. (2024) for details of the dataset). The pipeline detected upward and downward movements independently for each wrist. Human annotations specified the direction of movement (UP/DOWN) and the hand involved (Right, Left, or Both). Table 1 presents the results of the comparison between manual and automatic annotations. `No overlap` denotes human annotations for which no overlapping automatic annotation was found. Where the annotations overlap, `Strict` denotes agreement in movement direction for all annotated hands, `Relaxed` agreement for one hand, and `Wrong` no agreement in direction. For single-hand human annotations, a machine annotation is classified as `Strict` only when it identifies the same direction for the same hand and detects no movement for the other wrist.

The validation videos comprise `Studio` and `Home` recordings, which differ in lighting, camera angle, and video quality. These recording conditions affect the pipeline's ability to detect movements visible to human annotators, as reflected in the higher No-overlap rate for `Home` recordings. Once No-overlap cases are excluded, however, the proportions of `Strict`, `Relaxed`, and `Wrong` classifications are comparable across the two conditions.

6 Case Study

6.1 Previous findings and research questions

Wilson (under review) investigated how English speakers conceptualise time through co-speech gesture using manual multimodal annotation, statistical analysis, and conceptual blending analysis. The study demonstrated that English temporal conceptualisation is realised through a multidirectional gestural system rather than being restricted to the canonical forward-backward or rightward-leftward directions. Gestural direction was shown to be motivated by meaning construction in discourse, gesture multifunctionality and "gestural syntax" – a concept proposed by Wilson –, with sagittal, lateral and vertical directions jointly contributing to temporal meaning. The quantitative analysis of 151 manually annotated future and past speech–gesture instances identified the vertical axis as a discriminative parameter within the grammatical constructions examined: *will*-future constructions were predominantly accompanied by upward movement or positioning, whereas present perfect and simple past constructions were predominantly accompanied by downward movement or positioning. Qualitative analysis further demonstrated that the communicative function of vertical movement can only be interpreted in relation to the surrounding gesture sequence and discourse context rather than as an isolated movement.

The present case study investigates whether these findings generalise beyond the manually analysed 151 instances and how automatic annotation can support or extend such research. Specifically, we ask three related questions.

1. Does the systematic engagement of the vertical axis remain observable when vertical hand movement is analysed automatically across a substantially larger dataset?

2. What does this case study reveal about the validity and usefulness of automatic gesture annotation for multimodal communication research?

3. What new linguistic insights become possible once gesture can be analysed computationally at corpus scale?

To address these questions, we ultimately engaged with more than 80,000 speech units representing *will*-future, present perfect and past perfect to explore hand movement accompanying these temporal constructions in English.

6.2 Quantitative analysis

The NewsScape 2016 corpus was queried for combinations of *will* followed by one of the following

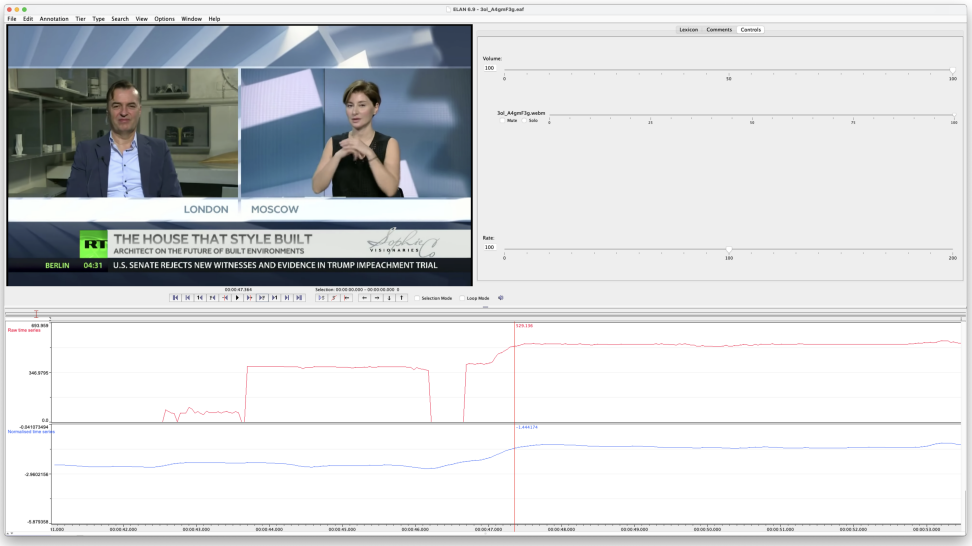


Figure 1: Time series of the vertical (y-)coordinate for the right wrist of the active speaker. The red time series (top panel) shows the raw OpenPose coordinates; zero values correspond to failures in detecting coordinates. The blue time series shows the post-processed (normalised and interpolated) version of the same signal, with less jitter and no missing values. The change in scale for the red line at time code 00:00:43.6 is due to a scene change. The gap at 00:00:46.2–00:00:46.6 is where OpenPose fails to predict the wrist's coordinates, visible as a drop in the raw signal but absent in the interpolated one.

Location	*N*	No overlap	Strict	Relaxed	Wrong
Home	329	152 (46.2%)	62 (18.8%)	66 (20.1%)	49 (14.9%)
Studio	99	12 (12.1%)	34 (34.3%)	33 (33.3%)	20 (20.2%)
Total	428	164 (38.3%)	96 (22.4%)	99 (23.1%)	69 (16.1%)

Table 1: Results of the annotation validation

verbs, which were inspired by Wilson (under review): *decrease*, *go down*, *increase*, *go up* (spatial: upwards and downwards), *stop*, *terminate*, *delete* (non-spatial: stopping), *read*, *swim* (non-spatial: neutral), and for sequences of *have*, *has* or *had* and the past participle of the same verbs. This yielded 7,926 hits, 839 of which were discarded because at least one of the words was not successfully aligned with the audio (see Uhrig (2022) for details). A further 84 instances with implausible lengths (less than 4 frames – approx. 133 ms or more than 90 frames – approx. 3 s) were removed. A further 2,914 instances were removed because no active speaker was detected. 2,408 hits were then removed because no keypoints were detected or no wrist had an average confidence score of at least 0.3. Out of the remaining 1,681, 579 showed no wrist movement above a small threshold used to filter resting hands and longer holds. Overall, more than 85% of query results were excluded, consistent with previous manual annotation of media data.

For the remaining 1,102 hits, wrist movements were counted separately for the auxiliary, content verb, and both; movements extending beyond the two-word span were excluded to minimise the influence of surrounding words.

The quantitative analysis demonstrates that the vertical axis is systematically engaged in all three grammatical constructions. Across the constructions, the auxiliary favours UP movement and the content verb DOWN movement. For movements spanning both positions, DOWN movements outnumber UP movements in *will*-future, UP movements slightly outnumber DOWN movements in present perfect, while past perfect is evenly divided (Table 5). These findings confirm that the engagement of the vertical axis identified by Wilson (under review) generalises to a substantially larger automatically annotated corpus, although the automatic UP/DOWN distributions show considerable overlap.

The overlap in the automatic UP/DOWN distributions could reflect limitations of the automatic

annotation pipeline, the complexity of naturally occurring multimodal communication, or the operationalisation of movement within predefined temporal windows.

To investigate these possibilities, we conducted a qualitative evaluation of 100 randomly selected clips from the automatically annotated dataset. Each clip was independently annotated manually in the Red Hen Rapid Annotator (Uhrig, 2022) by two expert human annotators and compared with the machine annotation. For every example we recorded observations concerning the quality of the automatic annotation together with its multimodal interpretation.

For a broader quantitative analysis, we queried the 102 verbs listed in Appendix A with *will*, *have*, *has* and *had*, and in the third-person singular present and past tense. The verbs were selected for their lack of direct temporal or spatial semantics. Random samples of 20,000 hits were extracted for *will*-future, present tense, present perfect and past tense, together with all 7,021 past-perfect hits. For each wrist movement, vertical amplitude was calculated as the difference between the y-coordinates of the last and first keypoints. Positive and negative values represent UP and DOWN movement, respectively (= signed amplitude). We also calculated absolute amplitude, i.e. the absolute value of the vertical distance. Peak velocity was calculated for combined lateral (x) and vertical (y) axes, and vertical axis, because the latter on its own does not indicate how fast the hand moves. Velocities for both types of measurements were calculated as the maximum change in coordinates between two consecutive wrist keypoints, multiplied by the frame rate.

The expanded quantitative analysis reveals differences between the three constructions. *Will*-future has the highest mean peak velocities for both UP and DOWN movements at the auxiliary, content-verb and combined positions (Table 3), and the largest absolute vertical amplitude at all three positions (Table 4). For absolute amplitude, comparisons between *will*-future and perfect constructions after movement-length normalisation were statistically significant in all cases except *will*-future versus past perfect at the auxiliary and content verb (Table 2). Absolute vertical amplitude increases from auxiliary to content verb in all three constructions, although this may be partly conditioned by our focus on movements within predefined temporal windows, and the longer duration of content verbs relative to auxiliaries (mean length auxiliaries: 169 ± 80 ms; content verbs: 196 ± 94 ms). UP peak velocities remain essentially stable in *will*-future, but decline in present perfect and past perfect (Table 3). This pattern is of interest given that DOWN peak velocities increase from auxiliary to content verb in all three constructions. Our analysis of mean signed amplitude confirmed that across the constructions the auxiliary favours UP movement and the content verb DOWN movement. The differences are statistically significant in all three constructions (Table 6, the raw data in Appendix B).

6.3 What does the case study reveal?

Research Question 1: Does the vertical-axis hypothesis generalise?

The quantitative analysis confirms that the systematic engagement of the vertical axis identified by Wilson (under review) generalises to a substantially larger automatically annotated corpus.

This study's qualitative analysis further supports Wilson's (under review) interpretation of the communicative function of the vertical axis. Human annotators evaluated each temporal grammatical construction as a single multimodal unit rather than treating the auxiliary and verb independently. Consequently, an upward movement or upward hand position associated with the *will* + infinitive construction was counted as positive evidence irrespective of whether it occurred on the auxiliary itself or began slightly before or ended shortly after the construction while remaining temporally associated with it. The latter cases primarily reflected the natural timing of gesture production, including instances where the articulatory onset of the gesture preceded the acoustic onset represented by subtitle alignment. Applying these criteria, 32 of 40 *will*-future constructions (80%) displayed the predicted upward pattern.

The same principles were applied to *has/have/had* + past participle constructions. Downward movement or downward positioning was counted irrespective of whether it occurred on the auxiliary itself (e.g. hands positioned down on has), on the participle while clearly belonging to the grammatical construction (e.g. left hand down on gone down; left hand up and down on increased), or began slightly before or ended shortly after the construction owing to the natural temporal coordination between gesture and speech. Applying these criteria, 30 of 39 constructions

Tenses	Position	Welch's t	p_{Bonf}
Will-future vs Present perfect	Auxiliary	3.2	0.009
	Content verb	4.4	5.54×10^{-5}
	Both	5.3	6.99×10^{-7}
Will-future vs Past perfect	Auxiliary	1.6	0.709
	Content verb	2.7	0.042
	Both	4.7	2.31×10^{-5}

Table 2: Welch's t-statistics on mean length-normalized vertical absolute amplitudes for *will*-future versus the two perfect constructions by verb segment; p-values are Bonferroni-corrected for six comparisons.

	Will-future		Present perfect		Past perfect	
Position	x, y	y only	x, y	y only	x, y	y only
Auxiliary – UP	1.59 ± 1.33	1.24 ± 1.02	1.35 ± 1.12	0.97 ± 0.86	1.48 ± 1.08	1.02 ± 0.77
Auxiliary – DOWN	1.33 ± 1.06	0.94 ± 0.84	1.24 ± 1.20	0.85 ± 0.72	1.13 ± 0.92	0.82 ± 0.68
Content verb – UP	1.58 ± 1.27	1.20 ± 1.12	1.25 ± 1.09	0.91 ± 0.76	1.23 ± 1.04	0.83 ± 0.68
Content verb – DOWN	1.45 ± 1.22	1.10 ± 1.02	1.36 ± 1.23	1.02 ± 1.03	1.38 ± 1.01	0.98 ± 0.72
Both – UP	1.97 ± 1.61	1.54 ± 1.35	1.87 ± 1.84	1.48 ± 2.11	1.96 ± 2.78	1.46 ± 2.55
Both – DOWN	2.07 ± 1.52	1.63 ± 1.37	1.70 ± 1.33	1.30 ± 1.44	1.76 ± 1.42	1.27 ± 1.02

Table 3: Mean peak velocity $\pm$ one standard deviation by position and construction, computed from the x and y (2D) coordinates and from the y coordinates only

(77%) displayed the predicted downward pattern.

The remaining (negative) examples demonstrate why temporal meaning cannot always be recovered from movement counts alone. Individual movements simultaneously fulfil multiple communicative functions, contributing not only to temporal reference but also to event depiction, discourse organisation, stance and interaction. Moreover, temporal interpretation frequently depends on the unfolding gesture sequence rather than on an isolated movement. The qualitative analysis therefore reinforces broader theoretical claim from Wilson (under review) that gestural direction is determined by multimodal meaning construction, gesture multifunctionality (see e.g. Kok et al., 2016) and gestural syntax rather than by fixed one-to-one correspondences between individual movements and linguistic forms (on gesture-speech timing see e.g. ter Bekke et al., 2024; Fritz et al., 2021; and cited literature).

Research Question 2: What does the case study reveal about automatic gesture annotation?

The comparison between machine and human annotation demonstrates the sufficient reliability of the automatic pipeline for large-scale analysis. Of the 100 annotated clips, 15 had to be discarded and excluded from the evaluation either because of errors in the subtitles or because of repetitions in the dataset. Of the remaining 85 instances, 35 (41.2%) showed complete agreement between machine and human annotation, while a further 32 (37.6%) showed partial agreement. Thus, 78.8% of the evaluated examples contained information produced by the automatic pipeline that was fully or partially consistent with expert manual annotation.

Many disagreements arose from factors external to the movement-detection algorithm itself, including partial hand visibility, occlusion by objects or television graphics, subtitle and forced-alignment errors, and camera movement. These limitations are characteristic of naturalistic broadcast data rather than of the computational method itself.

A further source of divergence between automatic and human annotation is that the former tracks wrist trajectories, whereas human annotators interpret the hand as a whole. The wrist and the fingers or palm may move or orient in different directions, producing different directional classifications of the same gesture.

Other discrepancies reflect fundamental properties of human communication. The automatic pipeline detects movement trajectories, whereas human annotators interpret communicative function. Every upward movement necessarily gives rise to a subsequent downward movement, and vice versa. Consequently, the movement detected automatically is not always the movement interpreted by human annotators as temporally foregrounded.

Position	*Will*-future	Present perfect	Past perfect
Auxiliary	.152 ± .183	.131 ± .158	.137 ± .141
Content verb	.190 ± .223	.160 ± .198	.163 ± .217
Both	.361 ± .355	.315 ± .330	.290 ± .306

Table 4: Mean absolute amplitude ± one standard deviation by position

Position/direction	*Will*-future	Present perfect	Past perfect
Auxiliary – UP / Total	482 / 826 (58.4%)	403 / 645 (62.5%)	123 / 183 (67.2%)
Content verb – DOWN / Total	884 / 1,522 (58.1%)	783 / 1,385 (56.5%)	228 / 393 (58.0%)
Both – UP / Total	909 / 2023 (44.5%)	814 / 1555 (52.3%)	230 / 460 (50.0%)
Both – DOWN / Total	1114 / 2023 (55.5%)	741 / 1555 (47.7%)	230 / 460 (50.0%)

Table 5: The fraction of UP or DOWN movements for the type of the verb by tense

Similarly, gestures frequently begin before or continue after the associated auxiliary, so the relevant movement may fall outside the grammatical construction selected for automatic analysis.

We experimented with temporal buffers around the target construction to capture movements beginning before or ending after the auxiliary. Although this approach successfully retrieved additional temporally relevant movements, it also admitted numerous unrelated gestures belonging to neighbouring discourse segments, thereby reducing the specificity of the automatic annotation. Increasing the temporal window therefore does not, by itself, solve the problem, but can be useful to improve recall (i.e. how many gestures are found) at the cost of precision (i.e. the rate of relevant gestures among those retrieved by the machine).

The qualitative evaluation additionally revealed several examples in which the automatic pipeline detected genuine movements that had initially been overlooked by human annotators. Subsequent detailed inspection confirmed that these movements were present in the video. Automatic annotation therefore does not simply reproduce manual analysis but can complement it by identifying subtle movements that may be missed because of attention limitations or annotation fatigue.

Research Question 3: New linguistic insights through automatic annotation at corpus scale The case study demonstrates that the engagement of the vertical axis identified through manual multimodal analysis generalises to a substantially larger automatically annotated corpus. *Will*-future has the largest absolute vertical amplitude across the auxiliary, content-verb and combined positions, and also the highest mean peak velocity for both upward and downward movements. This convergence is consistent with research linking gestural prominence to movement magnitude and other kinematic properties (Ferré (2018); Pouw and Dixon (2019)). In our data, the convergence of amplitude and peak velocity therefore suggests that the greater amplitude associated with *will*-future reflects more than simple spatial displacement and may index greater kinematic prominence with which the future is introduced into the multimodal communication (cf. Horst et al. (2014) on movement quality and communicative effort).

The construction-internal peak-velocity profiles show a further descriptive difference between *will*-future and the perfect constructions. DOWN peak velocity increases from auxiliary to content verb in all three constructions, while UP peak velocity remains essentially stable in *will*-future but declines in present perfect and past perfect (see Table 3). This pattern raises the possibility that *will*-future maintains a high-velocity upward component across the construction; however, the substantial variation within the data means that establishing whether this reflects genuine construction-internal stability requires more detailed quantitative and qualitative analysis.

7 Discussion and Conclusions

As we have shown, the case study demonstrates that automatic annotation is most valuable not as a replacement for expert multimodal analysis but as a complementary research instrument. The computational pipeline makes it possible to analyse thousands rather than hundreds of gesture instances, substantially reduces annotation effort, provides objective and reproducible movement measurements, and can identify genuine movements overlooked during manual annotation. Human analy-

Case	N (Aux)	N (Content)	Mean (Aux)	Mean (Content)	Welch t	Welch p
Will-future	826	1522	0.03 ± 0.24	-0.032 ± 0.29	5.7	1×10^{-8}
Present perfect	645	1385	0.035 ± 0.2	-0.038 ± 0.25	7.0	4×10^{-12}
Past perfect	526	1127	0.058 ± 0.18	-0.068 ± 0.27	11.2	6×10^{-27}

Table 6: Welch's t-test results by case

sis remains essential for interpreting communicative meaning within the unfolding multimodal discourse, whereas automatic annotation excels at detecting, measuring and retrieving gesture at scale. Together, the two approaches provide a scalable methodology for multimodal communication research that neither could achieve independently. If a sufficiently large dataset is available, the noise inherent in our system becomes manageable so that patterns can be discerned from fully automatically annotated data without human intervention in the annotation process, although we recommend always manually verifying a sample to confirm the patterns identified by the automatic analysis. Based on our analysis of a relatively large dataset, we can confirm the hypothesis from (Wilson, under review) that the vertical axis plays a role in the expression of time, while the combined automatic and qualitative analyses support the finding that *will*-future favours upward movement.

Linguistic contribution. Our case study demonstrates how automatic annotation can generate new linguistic questions that would not be practical to investigate through manual annotation alone. The ability to quantify properties such as movement amplitude and velocity across large samples makes it possible to investigate how gesture kinematics and multimodal prominence relate to the contrasting conceptualisation of future and past in speech–gesture communication. This extends previous work on multimodal prominence and the role of gesture in conceptualisation and information structuring into the domain of temporal conceptualisation at corpus scale (on multimodal prominence see Krahmer and Swerts, 2007; Swerts and Krahmer, 2008; Kita, 2000; Kita et al., 2017; Wilson, 2020 and the cited literature).

Methodological contribution. The case study demonstrates that computational pipelines and qualitative multimodal analysis answer complementary research questions. Automatic annotation captures movement trajectories at corpus scale reliably enough, whereas qualitative analysis establishes their communicative significance. Rather than replacing expert interpretation, computational annotation enables new forms of multimodal inquiry by combining large-scale quantitative analysis with detailed qualitative investigation.

8 Limitations

Automatic annotation and human multimodal analysis differ in what they capture. The automatic pipeline tracks the trajectories of individual wrists within predefined temporal windows, whereas human annotators perceive the hand as a whole, including the palm and fingers, and can interpret the movements of both hands together. Human annotators can also identify which movement is perceptually foregrounded and interpret it in relation to preceding and subsequent gesture and speech. The automatic analysis cannot make these contextual distinctions and restricts analysis to the selected temporal window, even though relevant gestures may begin before or continue beyond it.

The pipeline is also subject to technical limitations arising from keypoint extraction and normalisation. Partial hand visibility, occlusion and recording quality can impair keypoint detection, while camera movement creates additional problems for normalisation. In dynamic scenes, substantial camera movement can bias the reference coordinates used for normalisation, leading to shifted body coordinates. Frame-by-frame coordinate subtraction could reduce this problem, but may introduce apparent wrist movement when the wrist remains stationary while the speaker's head moves. We therefore use the normalisation procedure that proved most robust across the naturalistic broadcast data, while recognising that no approach eliminates these sources of noise completely.

Despite these limitations, the pipeline provides a scalable and reproducible means of detecting and measuring gesture across datasets that would be impractical to analyse manually, while also identifying movements that human annotators may overlook because of attention limitations or annotation fatigue.

References

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gaëlle Ferré. 2018. Gesture/speech integration in the perception of prosodic emphasis. In *Speech Prosody 2018*, pages 35–39.

Isabella Fritz, Sotaro Kita, Jeannette Littlemore, and Andrea Krott. 2021. Multimodal language processing: How preceding discourse constrains gesture interpretation and affects gesture integration when gestures do not synchronise with semantic affiliates. *Journal of Memory and Language*, 117:104191.

Zheyan Gao, Jinyan Chen, Yuxin Liu, Yucheng Jin, and Dingxiaofei Tian. 2025. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. *Artificial Intelligence Review*, 58(3):68.

Andrew Hardie. 2012. Cqpweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Dorothea Horst, Franziska Boll, Christina Schmitt, and Cornelia Müller. 2014. Gesture as interactive expressive movement: inter-affectivity in face-to-face communication. *Body–language–communication. An international handbook on multimodality in human interaction*, 2:2112–2125.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*.

Sotaro Kita. 2000. *How representational gestures help speaking*, page 162–185. Language Culture and Cognition. Cambridge University Press.

Sotaro Kita, Martha W. Alibali, and Mingyuan Chu. 2017. How do gestures influence thinking and speaking? the gesture-for-conceptualization hypothesis. *Psychological Review*, 124(3):245–266.

Kasper Kok, Kirsten Bergmann, Alan Cienki, and Stefan Kopp. 2016. Mapping out the multifunctionality of speakers' gestures. *Gesture*, 15(1):37–59.

Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. 2021. How to design a three-stage architecture for audio-visual active speaker detection in the wild. *arXiv preprint arXiv:2106.03932*.

Emiel J. Krahmer and Marc G. J. Swerts. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414.

Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14.

Mart Lubbers and Francisco Torreira. 2013-2026. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.71.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.

Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. 2022. Learning long-term spatial-temporal graphs for active speaker detection. In *European Conference on Computer Vision*, pages 371–387. Springer.

Siti Nur Aisyah Mohd Robi, Muhammad Atiff Zakwan Mohd Ariffin, Mohd Azri Mohd Izhar, Norulhusna Ahmad, and Hazilah Mad Kaidi. 2024. Active speaker detection using audio, visual, and depth modalities: A survey. *IEEE Access*, 12:96617–96634.

Robert Ochshorn and Max Hawkins. 2015. Gentle: a robust yet lenient forced aligner built on kaldi.

Wim Pouw and James A Dixon. 2019. Quantifying gesture-speech synchrony. In *The 6th Gesture and Speech in Interaction Conference*, pages 75–80. Universitaetsbibliothek Paderborn.

Marc G. J. Swerts and Emiel J. Krahmer. 2008. Facial expressions and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2):219–238.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 3927–3935.

Marlijn ter Bekke, Linda Drijvers, and Judith Holler. 2024. Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech. *Cognitive Science*, 48(1):e13407.

Peter Uhrig. 2022. *Large-scale Multimodal Corpus Linguistics: The Big Data Turn*. Habilitation, FAU Erlangen-Nürnberg.

Anna Wilson. 2020. It's time to do news again. *Zeitschrift für Anglistik und Amerikanistik*, 68(4):379–409.

Anna Wilson. under review. Rethinking gestural direction in english future conceptualisation: Corpus study of will-future and beyond. Submitted.

Anna Wilson, Irina Pavlova, Elinor Payne, Ilya Burenko, and Peter Uhrig. 2024. World futures through rt's eyes: multimodal dataset and interdisciplinary methodology. *Frontiers in Communication*, Volume 9 - 2024.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1556–1559, Genoa, Italy. European Language Resources Association (ELRA).

A The list of verbs

The list of verbs we used in section 6.2 for our quantitative analysis:

write, wear, wash, vote, use, urge, try, treat, trade, target, survive, surprise, succeed, struggle, spend, sing, show, shock, share, serve, sell, seek, search, satisfy, rush, risk, ring, respond, respect, request, report, reply, rent, remain, release, regard, record, recommend, read, quote, provide, prove, protect, prevent, prefer, prepare, practice, please, play, pick, persuade, pay, pack, oppose, operate, offer, offend, occur, obtain, notice, need, name, move, mix, miss, mind, mention, meet, mean, match, marry, manage, make, maintain, love, lose, load, live, listen, let, lend, leave, learn, laugh, know, kiss, judge, join, issue, involve, invite, invest, introduce, insist, inform, influence, infect, indicate, include, improve, imagine, ignore.

B Table

The table is the full version of Table 2 from the main text. It shows absolute vertical amplitudes ($|\Delta y|$ divided by movement length in frames, mean of per-movement values $\pm$ one standard deviation) for *will*-future versus the two perfect constructions, by verb segment.

C Post-processing of the OpenPose keypoints

C.1 Matching the Active Speaker to a Set of Keypoints

OpenPose predicts the coordinates of keypoints for all individuals present in a given frame, without reference to who is speaking. To identify which set of predicted keypoints corresponds to the active speaker, we use the bounding box predicted by TalkNet-ASD: for each frame, we identify candidate individuals whose head position lies within this bounding box. The nose keypoint serves as a proxy for head position; when OpenPose predicts the nose with low confidence, the average position of the eyes and ears is used in its place.[6]

OpenPose frequently detects several individuals in a single frame, and more than one detected individual's head position may fall within the TalkNet-ASD bounding box – for instance, when a second person appears in the background of the shot, or when two individuals are seated close together. Two checks are applied to select a single candidate in such cases.

First, a candidate is discarded if the ratio of their estimated head width to the bounding-box width falls outside the range $[1/3, 3]$. Head width is estimated from the inter-ear distance where both ears are reliably detected, or otherwise from the inter-eye distance scaled by a factor of 2.5; if neither is available, this check is not applied to the candidate. This filters out candidates whose apparent size is grossly inconsistent with the bounding box, such as a background head that happens to fall inside a box sized for a foreground speaker.

Second, among the remaining candidates, the individual whose head position is closest to the bounding-box centre is selected, with the horizontal and vertical components of this distance normalised by the bounding box's width and height, respectively. The frame is treated as ambiguous –

[6] A keypoint is considered reliable if its OpenPose confidence score is at least 0.2.

Tenses	Position	Mean $\lvert\Delta y\rvert$/length $\pm$ SD		Welch's t	p_{Bonf}
		Will-future	Tense in col. 1		
Will-future vs Present perfect	Auxiliary	0.028 ± 0.029	0.024 ± 0.018	3.2	0.009
	Content verb	0.029 ± 0.025	0.025 ± 0.022	4.4	5.54×10^{-5}
	Both	0.038 ± 0.031	0.033 ± 0.029	5.3	6.99×10^{-7}
Will-future vs Past perfect	Auxiliary	0.028 ± 0.029	0.025 ± 0.020	1.6	0.709
	Content verb	0.029 ± 0.025	0.025 ± 0.025	2.7	0.042
	Both	0.038 ± 0.031	0.032 ± 0.027	4.7	2.31×10^{-5}

Table 7: p-values are Bonferroni-corrected for six comparisons.

and no individual is selected – if the second-closest candidate's distance is less than 1.5 times the closest candidate's distance, i.e. if the two nearest candidates are not clearly separated in their proximity to the centre.

This pipeline is subject to two distinct failure modes, which are treated separately. The first concerns OpenPose's keypoint predictions themselves: even for a correctly identified speaker, OpenPose may fail to predict a given keypoint. This typically occurs under two conditions: when a body part is moving at high speed, resulting in a blurred region rather than a clearly resolvable shape, and when a body part is occluded by another object or person. Occlusion is prevalent in the videos analysed here; a hand may be occluded by another hand or by a lower-third graphic (see Figure 1). In such cases, OpenPose returns a coordinate value of zero. These values are treated as missing data and are subsequently interpolated (Section C.2).

The second failure mode concerns the matching procedure itself: either no detected individual's head position falls within the active speaker's bounding box, or the frame is ambiguous in the sense described above. In both cases, a NaN (Not-a-Number) value is stored in place of a coordinate. NaN values are subject to the same post-processing steps as zero-valued coordinates.

A further complication concerns the scale of the predicted coordinates. OpenPose reports keypoint locations in pixel values.[7] As pixel coordinates are a function of camera distance and angle, the same body part may be represented at substantially different scales across scenes and tracks, even for the same speaker. This variability limits the comparability of keypoint coordinates across a video and complicates both manual and automated analysis. The normalisation procedure described in Section C.4 is designed to address this issue.

[7]These predictions are not necessarily integers and occasionally contain small negative values, which we treat as an artefact of the prediction process.

C.2 Interpolation and Extrapolation

Missing values arising from either of the failure modes described above are addressed through interpolation and extrapolation, applied separately to each track.

Interpolation estimates missing values that occur between two valid measurements. Where a keypoint coordinate is missing but is bounded on both sides by valid (non-zero) values, the missing values are estimated by linear interpolation between these two points.

Extrapolation addresses missing values that occur at the beginning or end of a track, where no valid measurement exists on one side. In these cases, the nearest valid value is substituted for the entire run of leading or trailing zeros.

C.3 Smoothing

In addition to missing values, OpenPose keypoint predictions are subject to jitter: small frame-to-frame fluctuations in coordinate values that do not correspond to actual movement of the underlying body part.

To reduce jitter, each coordinate value is replaced with a weighted average of values within a sliding temporal window. For a given frame, the window comprises the N preceding values, the current value, and the N subsequent values, yielding a total of $2N + 1$ points. We use $N = 5$ (11 points), a value selected empirically as offering the best trade-off between removing jitter and preserving small but genuine movements.

Weights are assigned according to the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)$, evaluated at $2N+1$ points equally spaced on the interval $[-1, 1]$. This kernel assigns maximal weight to the central frame, with weight decreasing for frames further from the centre of the window. Figure 2 presents an

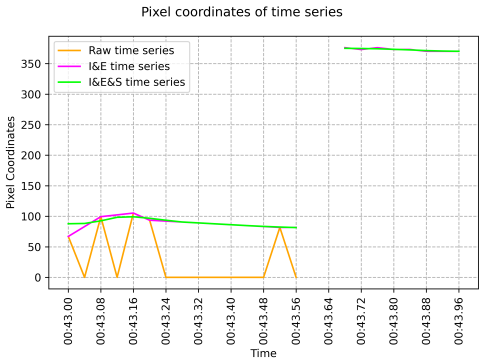


Figure 2: Raw, interpolated, and normalized and interpolated time series

example of a time series before and after smoothing.

C.4 Normalisation

Normalisation addresses the scale dependency introduced by the use of raw pixel coordinates (Section C.2). The procedure re-expresses keypoint coordinates relative to a reference point and a normalisation unit, both computed per track, yielding a coordinate space that is independent of the scale and position of the speaker within the frame.

As OpenPose predicts separate coordinate sets for the body, face, and hands, a reference point and normalisation unit are defined independently for each.

For the body, the nose coordinates serve as the reference point and are subtracted from all other body coordinates in every frame. The normalisation unit is the track-averaged distance between the nose (keypoint 0) and the neck (keypoint 1) in the BODY 25 annotation scheme. A vertical distance was chosen deliberately as the normalisation unit: TV and YouTube data frequently include scenes shot from the side, in which horizontal distances (e.g. shoulder width) are subject to foreshortening and therefore appear smaller than their true extent. Vertical distances are comparatively robust to this effect, as camera perspectives that substantially distort vertical measurements are considerably less common.

For the face, the tip of the nose serves as the reference point, and the normalisation unit is the track-averaged distance between the nasal bridge (keypoint 27) and the tip of the nose (keypoint 30) in the face keypoint annotation scheme.

Hand keypoints are not normalised, as the quality of OpenPose's predictions for hand keypoints is comparatively low, limiting the benefit of applying this procedure to them.

Normalisation of the body and face keypoints allows continuous time series to be produced across track and scene changes, even where the apparent scale of the speaker changes substantially, whether due to a change in shot for the same speaker or a change between speakers within the same scene.

Figure 2 illustrates the cumulative effect of interpolation and smoothing on an example time series. The orange curve represents the coordinates as originally predicted by OpenPose. The magenta curve shows the result of interpolation; extrapolation was not required in this example, as the original series did not contain leading or trailing zeros. The lime curve shows the result of smoothing applied to the interpolated (magenta) series.

Smoothing and normalisation together reduce the number of erroneously detected gestures, both in automatic annotation and in manual inspection of OpenPose-derived time series. This effect is illustrated in Figure 1, which compares raw values (top panel) with interpolated, extrapolated, smoothed, and normalised values (bottom panel) for the vertical coordinate of the left wrist. In the bottom panel, all coordinates are expressed on a common scale, without jitter or missing values.

Same Motions, Different Winners: The Gesture-Generation Leaderboard and Its Human Agreement Are Configuration-Tunable

Deep Gandhi[*]
Transformer Lab

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Abstract

Fréchet Gesture Distance (FGD) is the standard automatic metric for co-speech gesture generation, the models that synthesize the hand and body motion accompanying speech. Human perceptual studies remain the gold standard, and the GENEA challenges rank systems by them, but such studies are slow and costly, so papers routinely compare systems by FGD instead. FGD is a Fréchet distance in the latent space of a learned pose autoencoder, so it inherits the many choices behind that encoder, most of which papers never report. We ask whether FGD is a reliable proxy for perceptual evaluation under these unreported choices. Holding the systems fixed, we vary the reference tool's own settings (analysis window, seed, feature versus raw space) and the corpus the feature extractor is trained on, on both tiers of the GENEA 2022 evaluation and on the GENEA 2023 challenge. First, the FGD ranking is not reproducible where it matters: the "best" system is one of two that a per-clip bootstrap shows are statistically tied, so which one wins is a configuration choice, though aggregate rank correlation stays moderate rather than collapsing. Second, and more consequential, FGD's agreement with human ratings is modest and itself tunable, moving by 0.18 to 0.24 Kendall τ across these undocumented choices: FGD can be made to look appreciably more or less human-valid without changing the systems. These effects replicate across two challenge years and three evaluation tiers. We conclude that FGD numbers are not comparable across papers unless the window, seed, scoring space, and feature-extractor training corpus are pinned and reported, and that FGD cannot resolve small differences between closely performing systems.

1 Introduction

Co-speech gesture generation synthesizes the hand and body motion that accompanies speech, from audio and text, to animate embodied conversational agents and virtual avatars (Nyatsanga et al., 2023). Human perceptual studies are the gold standard for evaluating such motion, and the GENEA challenges rank systems by large crowdsourced user studies (Yoon et al., 2022; Kucherenko et al., 2024, 2023). Because such studies are slow and costly, papers routinely compare systems with Fréchet Gesture Distance (Yoon et al., 2020) instead, the automatic metric the challenges report alongside their human evaluation and the one surveys treat as standard (Nyatsanga et al., 2023; Nagy et al., 2024). FGD is the gesture analogue of FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018): a Fréchet distance in the latent space of a pose autoencoder trained to reconstruct motion. Because it is a distance in a *learned* feature space, it inherits every choice made in building that space, such as the corpus the encoder saw, the motion window length, the pose representation, and the random seed. These choices are rarely reported, yet FGD numbers are compared across papers that made different choices.

We ask whether FGD can be a reliable proxy for perceptual evaluation when only these nuisance choices change. Holding the systems fixed, we vary two families of configuration on the GENEA 2022 challenge outputs and their released human-likeness ratings: (A) the reference tool's exposed settings (window length, seed, and feature versus raw scoring), and (B) the corpus the feature extractor is trained on. We measure the stability of the induced ranking and its agreement with the human ranking, and repeat everything on the second GENEA 2022 tier and on the independent GENEA 2023 challenge.

We find two things. The FGD ranking is not reproducible at the decision-relevant margin, though aggregate correlation stays moderate rather than collapsing. More consequentially, FGD's agreement with human judgment is modest and can be

[*]Corresponding author: deep@lab.cloud

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 72–77
October 29, 2026 ©2026 Association for Computational Linguistics

moved appreciably by these undocumented choices. Because the choices are neither standardized nor reported, two papers computing FGD with different but reasonable settings can report different winners and different apparent validity without any change to the systems: FGD is unreliable for resolving small differences between closely performing systems, and cross-paper FGD numbers are not comparable. Our contributions are (i) a controlled audit of FGD across two GENEA challenge years that varies both the tool's exposed settings and the feature-extractor training corpus, measuring ranking stability and human agreement together; (ii) evidence that the top system flips between statistically tied systems while aggregate correlation stays moderate; and (iii) the main result, that FGD's human agreement is a researcher-tunable quantity, with concrete reporting recommendations.

2 Related Work

Fragility of Fréchet-style metrics. Fréchet-style generative metrics are known to be fragile to exactly the design choices we vary. Studies of FID, FVD, and Fréchet Audio Distance document content bias (Ge et al., 2024), sensitivity to sample size and the Gaussian assumption (Bischoff et al., 2024; Bench et al., 2026), misranking relative to human or downstream judgment (Wu et al., 2025; Jiralerspong et al., 2023), and dependence on the feature extractor and the choice of two-sample test (Chung et al., 2025; Cai et al., 2025; Grossi et al., 2025). These establish that the fragility exists in principle; what is missing is a measurement of how far it moves a real set of rankings.

Evaluating gesture generation. In gesture generation, human perceptual studies remain the gold standard (Kucherenko et al., 2024, 2023), but they are slow and costly, which is why FGD is used as their automatic proxy. Several works question that proxy: the GENEA Leaderboard position paper (Nagy et al., 2024) argues qualitatively for a standardized living benchmark, "2D or not 2D" (Guichoux et al., 2024) shows pose-representation dimensionality shifts FGD values for a pair of systems, and gesture perception has been examined in immersive settings (Werner et al., 2025). Closest to our setting, Crnek et al. (2025) search 56 feature-extractor and distance combinations on GENEA 2022 and 2023 for ones that better correlate with human studies (best ≈ 0.81); they seek a better metric, whereas we hold the standard FGD fixed and measure how far its undocumented configuration moves the rankings that papers already report, across two challenge years.

Robustness of automatic leaderboards. Our framing connects to a growing concern with the robustness of automatic leaderboards, from benchmarks that prove fragile to superficial changes (Feng et al., 2025) to leaderboards whose integrity can be undermined through the evaluation pipeline (Naseh et al., 2025). FGD-based rankings carry a similar but entirely undocumented degree of freedom in the metric's own configuration. Prior work asks which extractor best correlates with human studies; to our knowledge we are the first to audit the *ranking-comparability* consequences of the standard FGD's undocumented configuration, measuring the induced ranking instability and the change in human agreement together.

3 Method and Setup

Metric. FGD fits Gaussians to the pooled feature vectors of a candidate system's motion and of natural motion, and returns their Fréchet distance (Yoon et al., 2020); the feature space is the latent of an `EmbeddingNet` pose autoencoder. We use the GENEA reference implementation (Kucherenko et al., 2024),[1] train the autoencoder for 10 epochs (Adam, learning rate $5\mathrm{e}{-4}$; the released code uses 50, which we also test), and score every system against natural motion. Our default extractor is trained on the pooled motion of all systems, natural included. The challenge paper describes its extractor as "trained on the challenge dataset"; the released code trains it on the natural-motion test clips. Axis B below includes that natural-only corpus, so both readings are covered. The unit is a *system* (one model's released outputs); we never retrain a generator.

What we vary. The manipulated variable is the FGD configuration, along two axes. **Axis A** (the tool's settings) is a grid over window length $\{30, 60, 90\}$ frames, seed $\{0, 1\}$, and scoring space (the 32-dimensional latent, "feat", versus mean-pooled raw pose, "raw"), giving 12 configurations. **Axis B** (the training corpus) holds window at 30 and feature space, and trains the autoencoder on five corpora drawn from the same pool: all systems, natural only, all synthetic, and two disjoint halves

[1] github.com/genea-workshop/genea_numerical_evaluations

of the systems. The community treats the extractor as neutral; this axis tests that directly.

What we measure. We rank systems by FGD and compare rankings with Kendall's τ_b: the number of system pairs ordered the same way in both rankings minus the number ordered differently, divided by the number of pairs, with a correction for ties. $\tau = 1$ means identical order, 0 no association. With 10 synthetic systems there are 45 pairs, so τ moves in steps of $2/45 \approx 0.044$, and $\tau = 0.867$ means three pairs are ordered differently. We report three quantities. *Stability* is τ between the rankings of two configurations, plus whether the top-ranked system changes (a top-1 flip). *Validity* is τ between a configuration's FGD ranking and the human ranking, where the human ranking orders systems by their median human-likeness rating. *Spread* is the range of validity across configurations. We target human-likeness because FGD is a motion-realism metric matched to that axis, whereas the challenges' appropriateness rating turns on speech-gesture alignment that FGD does not model. Natural motion is the reference, so it attains near-zero FGD and trivially wins; headline numbers therefore use the *synthetic-only* ranking (natural excluded), with the with-natural counterpart reported alongside. Feat and raw are different metrics, so feat-only stability is primary. We quantify uncertainty by bootstrapping over study participants (validity), over the 40 clips per system (the FGD values), and by retraining at a larger budget (to rule out undertraining).

Data. We use the GENEA Challenge 2022 (Yoon et al., 2022), whose evaluation is analysed in extended form by Kucherenko et al. (2024):[2] the upper-body tier has 11 systems (10 synthetic plus natural), 40 clips each, pose dimension 162, 26,785 windows at length 30; the full-body tier has 10 systems and pose dimension 174, used for replication. We add a second challenge year, GENEA 2023 (Kucherenko et al., 2023)[3] (15 systems including natural, 30 clips each), whose outputs are released only as joint-angle BVH, so we convert them to 3D joint coordinates (root-relative, 330 dimensions) with the challenge's own forward-kinematics tooling, verified by a near-zero natural-versus-natural FGD. Systems carry anonymized GENEA identifiers (for example USQ, USM). Ratings are the released per-participant scores (roughly 1000 to 1500 per system, upper-body). All motion and ratings are public Zenodo releases (footnotes 2 and 3); no generator was trained.

[2]Motion: zenodo.org/records/6973297; ratings: zenodo.org/records/6940057.

[3]Motion: zenodo.org/records/8146027; ratings: zenodo.org/records/8434117.

4 Results

Ranking stability (Axis A): the aggregate order is moderately stable, but the winner is not. Across the 12 upper-body configurations, synthetic-only pairwise τ has median 0.867 and minimum 0.733; within the feature space alone it is higher (0.911 / 0.822). Rankings are never identical but the aggregate correlation is moderate, not a collapse: a typical configuration pair inverts a median of 3 of the 45 synthetic pairs, the most discordant pair inverts 6. Including natural raises stability slightly (0.891 / 0.782). What breaks is the top of the table: USQ ranks first in 4 of 12 configurations and USM in the other 8, so two systems are each "best" depending only on window, seed, and scoring space, and the flip occurs within the feature space. At window 30 the winner flips with the seed alone, and this persists when the extractor is retrained at five times the budget (50 epochs), so it is not an undertraining artifact. A per-clip bootstrap (500 resamples over the 40 clips; window 30, feature space, seed 0) shows USQ and USM are *statistically tied*: FGD 3.09 (95% CI $[2.21, 4.56]$) and 3.70 (95% CI $[2.51, 5.91]$), gap CI $[-1.35, 3.11]$, positive in 71% of resamples. The top-1 flip is thus a tie-flip: FGD cannot separate the two systems yet assigns a definite rank that flips with configuration, which is the mechanism behind our recommendation to treat close differences as ties.

Validity: human agreement is modest and moves with configuration. We treat validity as exploratory. Against the human ranking, synthetic-only τ has median 0.422, so FGD captures under half of the pairwise order of human preference. The disagreement is specific: at a median-agreement configuration (window 60, seed 0, feature space) FGD's best synthetic system USM is humans' 7th of 11, FGD's worst system USN is humans' 5th, and the human favourite USQ is FGD's 3rd. Including the natural anchor (which FGD trivially places first) raises the median to 0.491, matching the $\tau \approx 0.5$ FGD–human agreement reported by Kucherenko et al. (2024); the lower synthetic-only figure is the harder natural-excluded test, not a discrepancy with prior work. This agreement is tun-

able: across the 12 configurations it ranges from 0.333 to 0.511 (spread 0.178, standard deviation 0.057; Table 1), which corresponds to roughly four of the 45 system pairs reordering. Window length is one axis, but its direction is confounded: shorter windows yield more windows (window 30 about three times window 90) and Fréchet distance is sample-size biased (Bischoff et al., 2024). Holding the per-system window count fixed, the apparent shorter-is-better trend disappears and reverses (matched τ 0.289, 0.467, 0.333 at windows 30/60/90), so we make no recommendation on window length; the point is that agreement moves with the window. Window 30 and window 90 at their natural window counts (seed 0, feature space) do differ under participant resampling (paired difference 0.133, 0 of 2000 resamples ≤ 0, one-sided $p < 0.0005$; the marginal CIs overlap because they retain between-participant variance that the paired difference cancels, and τ over ten systems is quantized in 0.044 steps). Since those two configurations also differ in window count, this shows the choice is consequential, not that shorter is better.

Axis B: the extractor's training corpus is a hidden axis. Training the autoencoder on five corpora from the same pool yields synthetic-only cross-corpus τ of median 0.800 and minimum 0.644, comparable to, if not larger than, the reordering from the exposed settings. The mid-table moves (USN shifts from rank 10 to 7 by corpus), though the top is stable on this tier. Per-corpus human agreement ranges from 0.378 (all, natural, or all synthetic) to 0.556 (one arbitrary half), standard deviation 0.069: an undocumented, unmotivated choice moves apparent validity. A size-matched control (every corpus subsampled to the same window count before training the extractor) reproduces the reordering (synthetic-only τ 0.778 / 0.600) and most of the human-agreement spread (0.133 of 0.178), so the axis is primarily corpus content, not sample size. One confound remains: every corpus is drawn from the systems being scored, so we read this result as showing that the extractor is not neutral, without claiming which corpus is the right one.

Replication (full-body and GENEA 2023). Both effects reproduce on the full-body tier (Table 1). Axis A synthetic-only τ is 0.889 / 0.722, and the top system flips between FSA and FSH. Axis B cross-corpus τ is 0.806 / 0.667, and here the instability reaches the top: training on an arbitrary half flips the winner from FSA to FSH. Human agreement is again modest and tunable, with spread 0.222 on both axes. Which corpus agrees best differs by tier (full-body favours natural-only at $\tau = 0.611$), so the invariant is the large spread, not a consistent best corpus, which is itself evidence the choice is an unprincipled degree of freedom. The finding also holds on the GENEA 2023 challenge, a different year with different systems and a different skeleton: synthetic-only τ is 0.802 / 0.626 and the top synthetic system again flips (SJ versus SG), while agreement with humans is again modest (median 0.571) and configuration-dependent (spread 0.242).

	Stability τ		**Validity τ**	
Axis / tier	median	min	range	spread
A: settings (upper)	0.867	0.733	0.333–0.511	0.178
A: settings (full)	0.889	0.722	0.333–0.556	0.222
A: settings (2023)	0.802	0.626	0.451–0.692	0.242
B: corpus (upper)	0.800	0.644	0.378–0.556	0.178
B: corpus (full)	0.806	0.667	0.389–0.611	0.222

Table 1: Stability (synthetic-only pairwise Kendall τ between the FGD rankings of different configurations) and validity (synthetic-only Kendall τ between each configuration's FGD ranking and the human ranking, as a range over configurations and its spread). Rankings are never identical, the training corpus reorders the board comparably to the tool's own settings, and human agreement is modest and moves with configuration.

5 Discussion

No single FGD number is comparable across papers unless the window, seed, scoring space, and feature-extractor training corpus are all pinned and reported: two papers making different but reasonable choices can disagree on which system is best and on how human-valid FGD appears, without any difference in the systems. We recommend that FGD be reported with a fully disclosed configuration including the training corpus; that small FGD differences not be interpreted, since FGD cannot separate closely performing systems and even the top two here are statistically indistinguishable under clip resampling; and that cross-paper FGD comparisons be treated as unsound unless these choices are held identical.

6 Conclusion

FGD rankings of gesture-generation systems, and FGD's very agreement with human judgment, are

tunable by configuration choices that papers do not report. The ranking is not reproducible at the decision-relevant margin, where even the top two systems are statistically tied, and the metric's apparent validity can be moved by as much as 0.18 to 0.24 Kendall τ across tiers without changing the systems. These effects hold across two GENEA challenge years and three tiers. Comparable gesture evaluation requires pinning and reporting the full FGD configuration, including the feature-extractor training corpus, and treating close FGD differences as ties.

Limitations

The evidence is on the GENEA challenge datasets (two challenge years, three tiers); the direction should generalize to other corpora and motion formats, though magnitudes may not. Aggregate correlation is moderate (0.80 to 0.89), so the honest reading is that FGD is unreliable at the margin and for close systems, not that it is random. The training-corpus axis uses in-format subsets of the scored pool rather than an external corpus (the challenge training data is in a different motion format); a size-matched control rules out sample size, but the extractor still trains on the systems it scores, so a fully disjoint external corpus remains future work. The window-length trend does not survive matching sample size, so we claim no window-length effect; and the participant bootstrap covers rating noise, not FGD-side noise except where we bootstrap the FGD values. Kendall τ_b over ten systems is coarse (steps of $2/45$), which we reflect by reporting worst-case and typical discordance separately.

Availability

The code, runner, reference outputs, and data recipe are publicly available.[4] The audit uses only the public GENEA motion and rating releases cited in Section 3, which we do not redistribute.

Acknowledgements

The experiment setup and planning for this study were carried out with an earlier version of Primus, an automated research assistant developed by Transformer Lab.

[4] github.com/transformerlab-research/fgd-config-audit

Generative AI Usage

The experiment setup and planning used Primus, an LLM-based research assistant (see Acknowledgements). LLM tools (Claude Opus 4.8) were also used to format tables and to improve language clarity and grammar. The authors verified every reported number against the run outputs and take full responsibility for the content.

References

Ciaran Bench, Vivek Desai, Carlijn Roozemond, Ruben van Engen, and Spencer A. Thomas. 2026. Trustworthiness of Fréchet Inception Distance via Monte Carlo Dropout. *arXiv preprint arXiv:2601.21979*.

Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Glöckler, Lisa Haxel, Jaivardhan Kapoor, Janne K. Lappalainen, Jakob H. Macke, Guy Moss, Matthijs Pals, Felix Pei, Rachel Rapp, A. Erdem Sagtekin, Cornelius Schröder, Auguste Schulz, Zinovia Stefanidi, Shoji Toyota, and 2 others. 2024. A Practical Guide to Sample-based Statistical Distances for Evaluating Generative Models in Science. *Transactions on Machine Learning Research*, 2024. ArXiv:2403.12636.

Yuting Cai, Shaohuai Liu, Chao Tian, and Le Xie. 2025. Fréchet Power-Scenario Distance (FPD) for Smart Grids. *arXiv preprint arXiv:2505.08082*.

Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. 2025. KAD: No More FAD! Kernel Audio Distance for Audio Generation. *arXiv preprint arXiv:2502.15602*.

Karlo Crnek, Grega Močnik, and Matej Rojc. 2025. Advancing Objective Evaluation of Speech-Driven Gesture Generation for Embodied Conversational Agents. *International Journal of Human–Computer Interaction*. Published online July 2025.

Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. 2025. Visually Prompted Benchmarks Are Surprisingly Fragile. *arXiv preprint arXiv:2512.17875*.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. 2024. On the Content Bias in Fréchet Video Distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7288. IEEE. ArXiv:2404.12391.

Samuele Grossi, Marco Letizia, and Riccardo Torre. 2025. Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences. *Machine Learning: Science and Technology*, 6(1):015052. ArXiv:2409.16336.

Téo Guichoux, Laure Soulier, Nicolas Obin, and Catherine Pelachaud. 2024. 2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation? In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24)*, pages 22:1–22:4. Association for Computing Machinery. ArXiv:2409.10357.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30. Curran Associates, Inc. ArXiv:1706.08500.

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. 2023. Feature Likelihood Divergence: Evaluating the Generalization of Generative Models Using Samples. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pages 33095–33119. Curran Associates, Inc. ArXiv:2302.04440.

Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A Large-Scale Evaluation of Gesture Generation Models in Monadic and Dyadic Settings. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*, pages 792–801. Association for Computing Machinery. ArXiv:2308.12646.

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2024. Evaluating Gesture Generation in a Large-Scale Open Challenge: The GENEA Challenge 2022. *ACM Transactions on Graphics*, 43(3):32:1–32:28. ArXiv:2303.08737.

Rajmund Nagy, Hendric Voss, Youngwoo Yoon, Taras Kucherenko, Teodor Nikolov, Thanh Hoang-Minh, Rachel McDonnell, Stefan Kopp, Michael Neff, and Gustav Eje Henter. 2024. Towards a GENEA Leaderboard – an Extended, Living Benchmark for Evaluating and Advancing Conversational Motion Synthesis. *arXiv preprint arXiv:2410.06327*.

Ali Naseh, Anshuman Suri, Yuefeng Peng, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. 2025. Text-to-Image Models Leave Identifiable Signatures: Implications for Leaderboard Security. *arXiv preprint arXiv:2510.06525*.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum*, 42(2):569–596. ArXiv:2301.05339.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards Accurate Generative Models of Video: A New Metric and Challenges. *arXiv preprint arXiv:1812.01717*.

Axel Wiebe Werner, Jonas Beskow, and Anna Deichler. 2025. Gesture Evaluation in Virtual Reality. *arXiv preprint arXiv:2509.12816*.

Yuli Wu, Fucheng Liu, Rüveyda Yilmaz, Henning Konermann, Peter Walter, and Johannes Stegmaier. 2025. Pragmatic Note on FID for Retinal Image Synthesis. *arXiv preprint arXiv:2502.17160*.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics*, 39(6):222:1–222:16. ArXiv:2009.02119.

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22)*, pages 736–747. Association for Computing Machinery. ArXiv:2208.10441.

Behavioral Alignment and the Engagement for Non-Speaking Participants in Online Meetings

Trung Duc Anh Dang[*] University of Copenhagen

Zhiqing Peng [*] University of Copenhagen

Patrizia Paggio[†] University of Copenhagen
paggio@hum.ku.dk

Abstract

Engagement detection in meetings has focused mostly on speakers, who produce explicit prosodic and acoustic cues. The engagement of non-speaking participants is harder to read because it surfaces only in nonverbal behaviour. We predict how engaged a silent listener is, on a three-level scale (Engaged, Intermittent, Disengaged), in five-second windows of remote Zoom meetings from the GEstures and Head Movements in Language (GEHM) corpus. Each window is described by three feature families: the listener's own nonverbal behaviour, the listener's behavioural alignment with the others present, including the active speaker, and that speaker's prosody. Interpersonal coordination is an established correlate of rapport and involvement. We ask the narrower question of whether a compact window-level alignment measure adds predictive value over listener-only features when confident labels are scarce. Under leakage-free cross-validation, listener-speaker alignment is the most useful addition we test, and the speaker's terminal pitch slope is an interpretable correlate of listener engagement. The gains are modest and their bootstrap intervals include zero. Unsupervised analysis finds no engagement structure. Two findings are firmer: the task is effectively binary, actively engaged versus not, bounded by the scarcity of confident disengaged labels, and the pitch-slope correlate holds despite the small sample.

1 Introduction

Online meetings have become a standard means of communication across work and education due to the increasing need for remote collaboration, and engagement detection in this setting is an active research area, since engagement is a key indicator of meeting quality, productivity, and effectiveness (Frank et al., 2016). Research on engagement detection mainly targets the speakers, who produce explicit and quantifiable engagement signals, especially prosodic-acoustic features (Pellet-Rostaing et al., 2023). Feedback from listeners, conversely, is often conveyed through subtle and multifunctional non-verbal cues, making it challenging to interpret and classify their level of engagement (Pellet-Rostaing et al., 2023). Listener engagement is also an important factor shaping meeting dynamics and productivity, which motivates analyzing the engagement of non-speaking participants directly.

The unit of prediction is a five-second window in which a target participant is *silent* while someone else holds the floor, taken from the GEHM corpus of remote multi-party Zoom meetings (Paggio et al., 2024). For each window we predict how engaged the silent target appears, on the three-level scale used by our annotators: *Engaged*, *Intermittent*, and *Disengaged* (defined in §3.1). We adopt the behavioural definition of engagement proposed by Fredricks et al. (2004): visible participation and involvement in interaction, read from head pose, eye gaze, and facial expression (Kaur et al., 2018). The target produces no words in these windows, so engagement must be read from nonverbal behaviour and from the surrounding interaction.

Various behavioral features have been used for engagement detection and classification, including prosodic-acoustic cues, facial expressions, gestures, postures, head pose, and gaze behaviors (Oertel et al., 2011; Singh et al., 2023; Inoue et al., 2018). Recent engagement estimation models take conversational context into account, estimating a target participant's engagement by combining the participant's own audio-visual behaviour with cues from their conversational partners (Li et al., 2024; Müller et al., 2024). These methods depend on supervised engagement labels. They integrate the partner through learned audio-visual representations, not through an explicit measure of the be-

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 78–88
October 29, 2026 ©2026 Association for Computational Linguistics

haviour relation between the two.

In social signal processing and social neuroscience, interpersonal coordination is an established correlate of rapport, affiliation, and involvement (Bernieri et al., 1991; Delaherche et al., 2012; Ohayon and Gordon, 2025); we review this work in §2. Our question is therefore not whether coordination relates to engagement but whether it can be operationalised usefully under the constraints of this setting. Does a compact, window-level measure of how similarly a silent listener behaves to the others present add predictive value over listener-only features? The setting is unfavourable: the measure comes from off-the-shelf visual representations, the meetings are remote and lack the mutual-gaze and proxemic cues of co-present interaction, and confident labels are few. We treat this as a question about incremental predictive value in an ablation. Our analysis is correlational and confined to a single window. They cannot causally separate a listener's individual behaviour from their coordination with others; we return to this in the Limitations section.

First, we describe an exploratory clustering experiment in which we investigate whether engagement behaviour can be learnt in an unsupervised way. Then we report the results of several classification experiments in which engagement is first predicted in terms of three classes, and then based on a binary distinction, using different feature families.

Our contributions are: (1) **An explicit window-level alignment measure for silent listeners:** We quantify how much the listener's behavioural alignment with others adds over listener-only behaviour. In our ablation, listener-speaker alignment is the strongest addition, ahead of alignment with other listeners and of speaker-only features. The margins are within fold-to-fold variance and their bootstrap intervals include zero. The speaker's terminal pitch slope is an interpretable correlate of listener engagement. (2) **Label-scheme sensitivity analysis under label scarcity:** We test several binary reductions of the original three-class engagement scheme to examine whether the task difficulty comes from weak behavioural signal or from the label structure itself. This provides a basis for interpreting whether poor fine-grained performance reflects insufficient features or ambiguity in the engagement labels.[1]

[1]Code, feature-extraction pipeline, and annotation guidelines are available at https://github.com/ducanhdt/listener-meeting-engagement.

2 Related Work

Engagement in conversation and meetings. Engagement and involvement have often been detected through prosodic and movement cues (Gatica-Perez et al., 2005). Oertel et al. (2011) found that prosodic cues, together with head and body movement, correlated with the level of involvement in conversation. These cues are less available for non-speaking participants, whose engagement is expressed mainly through subtler nonverbal behaviour. Recent engagement estimation studies also use cues from conversational partners. The Dialogue-Aware Transformer framework (DAT) by Li et al. (2024), for instance, uses both the participant's behaviour and information from their conversational partner. These models learn participant–partner interactions from supervised audio-visual data, whereas we explicitly measure the listener's behavioural alignment with other meeting participants, including the speaker, against a small manually annotated dataset.

Listener cues and label subjectivity. For silent listeners, previous work has focused mainly on feedback behaviour. Inoue et al. (2018) recognize engagement from listener behaviours such as backchannels, laughter, nodding, and gaze, while also noting that engagement judgments vary across annotators. Related work in online learning also predicts engagement from visual cues such as facial expressions, head pose, posture, and eye movement, often in settings where learners are not speaking (Dewan et al., 2019; Kaur et al., 2018). We adopt similar visual descriptors for silent listeners, while facing the same central challenge: engagement labels are partly shaped by annotator uncertainty.

Classifying the silent participant. The closest precedent for our task is Oertel et al. (2015), who study silent participants in co-present group discussions. Raters assessed both the *listener role* a silent participant occupies and their *engagement level* from 15-second video clips, and the authors then classified those roles automatically from audio-visual cues. Our scheme keeps only the second of these dimensions: the three classes grade the degree of attentional involvement a listener displays in a five-second window (§3.1), without distinguishing whether that listener is the one currently addressed. Role is therefore an uncontrolled confound for our labels, since some of what an annotator reads as intermittent involvement may be

a listener who is simply not being addressed. The settings differ as well. Their discussions are co-present; ours are remote gallery-view Zoom meetings, where mutual gaze is not recoverable and role is correspondingly harder to infer.

Interpersonal coordination as a social signal. That interacting people coordinate their behaviour, and that the degree of coordination tracks the quality of the interaction, is established. Bernieri et al. (1991) showed that observer-rated movement synchrony predicts judgments of rapport. Delaherche et al. (2012) survey how synchrony has been operationalised across psychology, social signal processing, and clinical research; the measures differ in time scale and in what counts as similarity. The meta-analysis of Ohayon and Gordon (2025) finds a positive association between multimodal interpersonal synchrony and social outcomes such as affiliation and cooperation, with substantial heterogeneity across modalities and measurement choices. Wohltjen and Wheatley (2021) link coordination to attention directly: in dyadic conversation, moments of eye contact coincide with peaks of pupillary synchrony and are followed by its decline, so more synchrony does not simply mean more shared attention. In language technology the same idea appears as entrainment. Levitan et al. (2011) show that speakers entrain on acoustic-prosodic cues preceding backchannels, and that greater similarity accompanies better dialogue coordination and task success (Levitan et al., 2012). Müller et al. (2018) predict low rapport in small groups from non-verbal behaviour including synchronisation features. Closest to our data, Paggio et al. (2026) show that in the GEHM corpus speakers move their heads more in target intervals where verbal feedback is present.

Our position. We do not claim that coordination is a new indicator of engagement; the work above already supports it. What the literature does not settle is whether a compact, unlagged, window-level similarity, of the kind a small labelled sample can support, retains any of that signal for silent listeners in remote meetings. That is the question we test.

3 Methodology

We predict the engagement of a target participant in short windows in which that participant is *silent*. Within each meeting we divide every participant's video into five-second windows with a two-second overlap. For a given target we keep only the windows in which the target produces no speech while another participant holds the floor. Each such window is the unit of prediction. Engagement must therefore be read from the target's nonverbal behaviour and from the surrounding interaction, since the target participant by definition produces no words in the windows under consideration. Figure 1 summarises the full pipeline.

3.1 Corpus and Labels

We use the GEHM corpus (Paggio et al., 2024), collected by one of the present authors in an earlier project: twelve remote multi-party meetings (about eight hours, English) recorded over Zoom, so each participant appears head-and-shoulders in a separate gallery-view stream with a separate audio channel, and is mostly silent while listening to whoever currently holds the floor. The engagement annotation described below was produced for this study and is new.

Each participant's video is cut into five-second windows with a two-second overlap. We use the TextGrid speech alignments to remove windows in which the target speaks, including questions and short verbal responses, and drop windows where the target is not visible in at least 95% of frames (OpenPose detection; Cao et al., 2021) or faces substantially away from the screen, judged by average head orientation. Near-duplicate windows are clustered by CLIP (Contrastive Language–Image Pre-training; Radford et al., 2021) embedding, keeping one representative per cluster to reduce visual redundancy. Each retained window is paired with two randomly chosen co-present participants from the same timeframe, whose streams supply the reference signals for the listener $\leftrightarrow$ co-participant alignment. After filtering, 549 silent windows remain.

The active speaker for a window is the co-participant whose transcript overlaps it the most, identified for about 64% of windows; windows with no overlapping speech have no identified active speaker, and the speaker family is then imputed with the training-fold median (§3.3).

Engagement definitions. Annotators judged, for each five-second window, how engaged the silent participant appeared *in that window*, using the behavioural notion of engagement already mentioned in §1: visible participation and involvement, read off head pose, gaze direction, facial expression, and head and hand movement (Fredricks et al., 2004;

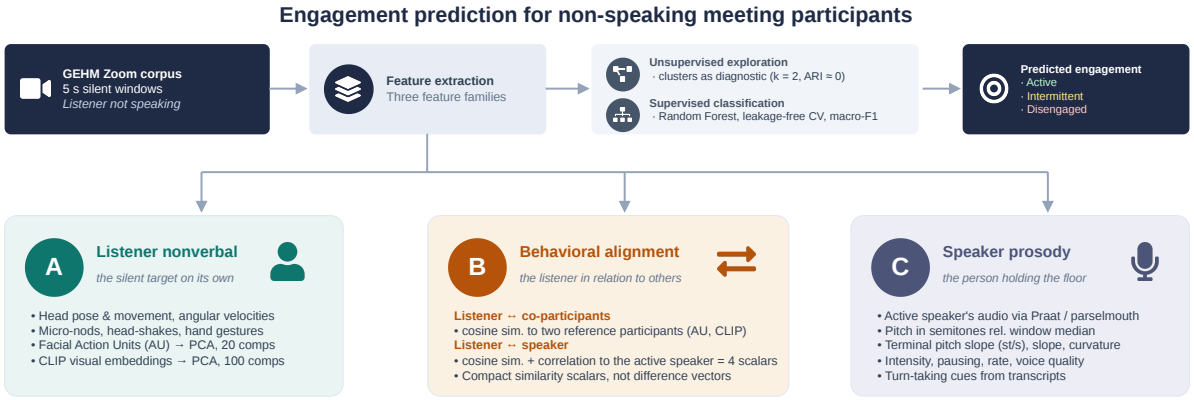


Figure 1: Overview of the approach. Each five-second window in which the target participant is silent is described by three feature families: the listener's own nonverbal behaviour (A), the behavioural alignment between the listener and others (B), and the active speaker's prosody (C). The same features feed an unsupervised exploration, where clustering is used only as a diagnostic, and the supervised classification that predicts the listener's engagement.

Kaur et al., 2018). The guideline defined the three levels as follows. *Engaged* (E): the target appears to be attending to the ongoing talk throughout the window, typically oriented towards the screen, with responsive behaviour such as nodding, gaze tracking the active speaker's tile, or facial reactions that follow the content. *Disengaged* (D): the target shows no sign of attending to the ongoing talk, for instance sustained gaze away from the screen, sustained absence of responsive movement, or visible involvement in an unrelated activity. *Intermittent* (I) was intended as the label for windows that could not confidently be placed in either of these two classes: the target shows mixed or weak cues, for example occasional distraction while still broadly following the speaker, so that the annotator cannot say with confidence whether the target is engaged or not. Annotators marked a window *unclear* when they could not make any judgment at all. The middle class is thus by design a region of annotator uncertainty rather than a distinct behavioural state, which bears on how the three-class results should be read (§4.4).

Annotation and agreement. Four non-expert annotators labelled 107 selected windows under this guideline. They saw only the target's video stream for the window, without audio and without the other participants' streams, so the judgment rested on the target's visible behaviour rather than on what was said or on how others reacted. One annotator's labels diverged from the other three (mean pairwise Cohen's $\kappa = 0.17$ (Cohen, 1960), far below the agreement among the other three) and were excluded. The exclusion was decided from the agreement matrix alone, before performing any experiment. The three retained annotators reach Fleiss' $\kappa = 0.42$ (Fleiss, 1971), moderate agreement. Engagement judgments from thin slices of silent video vary across annotators (Inoue et al., 2018), and most of our disagreement falls on the middle *Intermittent* class rather than on the *Engaged*/*Disengaged* contrast. Windows without a majority label, or marked *unclear* by most annotators, were removed, leaving 102 windows (46 unanimous, the rest by majority vote). Removing windows that lacked two reference participants gives the final 89. The supervised experiments use those 89 windows (44 Engaged, 28 Intermittent, 17 Disengaged); the unsupervised exploration (§4.1) uses all 549 silent windows.

Restricting the evaluation to windows with a clear majority label is a trade. It removes the windows where the label is least trustworthy, at the cost of evaluating on a smaller and easier sample, so the absolute scores we report are an upper bound on performance over all silent windows. The resulting 89-window sample is small and class-imbalanced, which is why we report against a majority-class floor and treat all gains as directional. All unsupervised steps, standardisation, PCA, and the speaker-feature imputation statistics, however, are fit on all 549 silent windows, not only the 89 labelled ones. We treat the three-class formulation as primary because it is what the annotators used. The moderate agreement on the middle class is why we also report a family of label reductions (§3.3, §4.4) to verify whether the difficulty lies in the behaviour or in the label scheme.

3.2 Feature Families

We represent every window with three families of features, organised so that the contribution of the *interaction* can be isolated from that of the listener alone.

Family A: listener nonverbal behaviour. The first family characterises the target in isolation, following the head-pose and head-movement descriptors used for engagement in mostly silent settings such as online learning (Gupta et al., 2016; Dewan et al., 2019). From per-frame body and face keypoints (Cao et al., 2021) and estimated head orientation, we derive head-pose statistics, angular velocities, and counts of micro-nods and head-shakes, together with hand-gesture cues detected from wrist and face keypoints (Lugaresi et al., 2019). We add per-frame facial Action Units (AU) (Baltrušaitis et al., 2018) and CLIP visual embeddings (Radford et al., 2021) of the listener's frames. The Action Units and CLIP embeddings are each reduced by principal component analysis to the number of components that preserves nearly all of their variance: 20 components for the Action Units and 100 for the CLIP embeddings, retaining roughly 100% and 97% of the variance, respectively. This family is the target-only baseline against which the interactional families are measured.

Family B: behavioral alignment. The second family measures how closely the listener's behaviour resembles others in the meeting, motivated by the long-standing finding that interpersonal entrainment and movement synchrony reflect rapport and affiliation (Levitan et al., 2011; Bernieri et al., 1991; Müller et al., 2018; Ohayon and Gordon, 2025). Among the many ways synchrony has been operationalised (Delaherche et al., 2012), ours is at the simple end: an unlagged, within-window similarity between representations of two people's visible behaviour, a choice driven by sample size rather than theory (see Limitations). We compute alignment along two distinct relations. *Listener* $\leftrightarrow$ *co-participants* alignment is the cosine similarity between the target and two randomly chosen co-present listeners, together with the similarity between those two, on the AU and CLIP channels, capturing whether the listener behaves in step with the others and stays oriented toward the meeting rather than looking away. *Listener* $\leftrightarrow$ *speaker* alignment is computed directly between the target and the single concurrently active speaker in the same window, on the visual (CLIP) and Action-Unit channels, via cosine similarity and correlation. We encode alignment as a small set of similarity *scalars* rather than as a high-dimensional element-wise difference vector, which keeps this family compact enough to be usable given the relatively small sample size we are working with.

Family C: speaker prosody. The third family characterises the person currently holding the floor, on the premise that the prosody of ongoing speech is part of the interactional context that shapes how engaged listeners appear: speakers provide prosodic cues, terminal pitch movement among them, that invite listener feedback (Gravano and Hirschberg, 2011). The active speaker for a window is the co-participant whose speech overlaps it the most. From that speaker's separated audio, we extract acoustic prosody features with Praat (Boersma and Weenink, 2021) and the Parselmouth library (Jadoul et al., 2018), expressed in semitones relative to the window median so that it is comparable across speakers: mean pitch and pitch variability, overall and terminal pitch slope, intensity, pausing, speaking rate, and voice quality. From the time-aligned transcripts, we also retain conversational turn-taking cues: how many participants are speaking in the window, how long the current speaker has already held the floor, how recently the listener last spoke, how much speech from others overlaps, and the window's position in the meeting. Because the speaker generally addresses the group rather than the specific silent listener, we read these cues as concurrent interactional *context* that co-varies with the listener's state, not as a signal directed at the listener.

3.3 Experimental Setup and Evaluation

Unsupervised exploration. Given the fact that reliable annotated labels of engagement only cover a small subset of the material, we first ask whether engagement can be recovered *without* them. This follows a semi-supervised intuition: listeners attending to the same speaker should behave similarly, so engagement states might surface as natural groups in the behavioural feature space, which a handful of labelled examples per group could then name. We cluster the windows with three standard algorithms, KMeans, a Gaussian Mixture Model (GMM), and Ward agglomerative clustering, on two feature configurations, the listener-only baseline and the baseline augmented with alignment,

and select the number of groups by silhouette width. We then compare the clusters with the gold labels through the Adjusted Rand Index. We treat clustering as a diagnostic of whether engagement is a salient axis of behavioural variation, not as a replacement for classification (§4.1).

Supervised classification. The supervised pipeline measures the contribution of each feature family with stratified five-fold cross-validation. All standardisation, PCA, and imputation are fit on the training folds only, so the protocol is leakage-free. In particular, speaker features that are missing when no speaker is identified are assigned the value of the training-fold median. Because the classes are imbalanced (44/28/17), we report macro-F_1 as the primary metric, with accuracy and a majority-class floor (always predicting the most frequent class) as references. The primary estimator is a Random Forest of 300 trees with balanced class weights, which handles the high-dimensional visual features well; we also report Logistic Regression, with balanced class weights and up to 2000 iterations, to confirm the direction of effects. We quantify the sampling uncertainty of each comparison with a paired bootstrap: 5000 resamples stratified by the true label, sampling windows with replacement within each class so every replicate keeps all classes and the original class balance, from which we report a 95% percentile interval for every macro-F_1 margin. The bootstrap holds the fold models fixed and reuses their out-of-fold predictions, so it estimates the variability due to which windows are labelled rather than model refitting; it is paired, meaning the same resampled windows score every configuration, so a margin's interval is not inflated by between-configuration differences, and the one-sided p is the resampled probability that a margin is at most zero. All models use scikit-learn (Pedregosa et al., 2011) under a fixed random seed (42), including the clustering algorithms used in the unsupervised exploration. To separate engagement signal from annotation noise, the same protocol is applied to several binary reductions of the original three-class label scheme as a sensitivity analysis.

4 Results and Discussion

4.1 Initial Exploration: Does Engagement Cluster?

Our starting point, following the original formulation of this task, was semi-supervised (§3.3): if listeners attending to the same speaker exhibit similar behavioural patterns, different engagement states should form distinguishable clusters in the feature space, which can then be interpreted using a small number of labelled examples. This is particularly useful in settings where high-confidence labels are scarce. We therefore cluster the windows before training any classifier and ask whether the groups coincide with the three engagement classes; the labels are used only afterwards, to check the correspondence.

Two findings emerge. First, the feature space prefers *two* groups rather than three: silhouette width peaks at $k = 2$ (KMeans 0.098, against 0.030 at the annotated $k = 3$), cross-checked with the GMM's Bayesian and Akaike Information Criteria (BIC/AIC) and a Ward dendrogram. Second, the groups do not correspond to engagement: forcing $k = 3$, the Adjusted Rand Index between clusters and gold labels is near zero (-0.01 to -0.03), and adding alignment features does not change this.

This second result holds across the other two clusterers, so it is not an artefact of one algorithm: Gaussian mixture is degenerate on this sample, collapsing the labelled windows into a single component with a silhouette curve that coincides with KMeans at every k, and Ward agglomerative clustering agrees on the substance, with weak structure, a marginal $k = 3$ at silhouette 0.086, and no label recovery. Instead, the clusters separate different behavioural styles of listening, such as stillness versus visible movement, which cut across engagement levels. Clustering thus recovers real structure, but that structure is behavioural *style*, not engagement, so the semi-supervised plan would mostly propagate style. We read this as motivating rather than negative: engagement is not the dominant axis of behavioural variation, which is why the rest of the paper predicts it with supervised models. The preference for two groups also anticipates §4.4, where the distinction the data supports is shown to be binary rather than the annotators' three classes.

4.2 The Contribution of Alignment Features

Among the families we test, *alignment* features, which describe the listener only in relation to oth-

ers, give the largest improvement over a listener-only model. Table 1 reports the supervised factorial ablation: each feature family is stacked on the listener-only baseline under the leakage-free protocol described in §3.3. The baseline already clears the majority-class macro-F_1 baseline (macro-F_1 0.341 vs 0.220). The best addition is listener ↔ speaker alignment, which raises macro-F_1 to 0.378, the highest in the table, and lifts accuracy to 0.505. It exceeds listener ↔ co-participant alignment (0.326), the speaker's own prosody (0.306), and the listener-only baseline itself. The ordering matches what the synchrony literature (§2) predicts: coordination with the person holding the floor carries more social signal than coordination with co-present non-speakers. It is an ordering of point estimates whose margins, quantified below, sit within fold-to-fold variance.

What gain there is comes from measuring the relation, not from giving the model more features. Listener ↔ speaker alignment produces the largest improvement while adding only four dimensions to the baseline ($166 \rightarrow 170$), namely the cosine similarity and correlation between the listener and the speaker on the visual (CLIP) and Action-Unit channels, whereas folding in the speaker's raw prosody adds thirty-seven dimensions ($166 \rightarrow 203$) and *lowers* macro-F_1 to 0.306. The blocks do not stack, either: the compact speaker-alignment on its own beats every richer combination (co-participant alignment + prosody 0.364, all three 0.337), because at $n = 89$ the extra dimensions cost more than they add. A four-number summary of how two people co-behave is the single best configuration we find, and that parsimony is what makes the relation usable at this sample size.

Accuracy changes little because the task is imbalanced. A classifier that always predicts the majority class scores accuracy 0.494 but macro-F_1 only 0.220; our balanced-weight models trade raw agreement for minority-class recall, so the listener-only baseline scores below that accuracy floor (0.459) even as its macro-F_1 rises, and adding speaker alignment reaches just 0.505. On a $44/28/17$ split, gains in minority-class recall rather than raw agreement are exactly what we want, and the reason we read macro-F_1 as primary.

Across the five cross-validation folds, macro-F_1 already varies by about 0.09 simply from how the data happens to be split, so the differences we describe here are directional rather than significant: most fall within that natural fold-to-fold range.

Configuration	Macro-F_1	Acc.	Dim
Majority baseline	0.220	0.494	—
Listener only (baseline)	0.341	0.459	166
+ Co-participant align.	0.326	0.482	172
+ Speaker align.	**0.378**	**0.505**	170
+ Speaker prosody	0.306	0.483	203
+ Co-participant align. + prosody	0.364	0.517	209
+ All families	0.337	0.494	213

Table 1: Supervised factorial ablation on the three-class target, stacking interactional feature families on the listener-only baseline (the nonverbal family, A). Co-participant alignment and speaker alignment are the two relations of the alignment family (B); speaker prosody is family C. Cross-validated Random Forest macro-F_1 and accuracy (standard deviation across folds ≈ 0.09 throughout); *Dim* is the feature dimensionality.

A paired bootstrap over the held-out predictions (§3.3) makes this concrete: the 95% interval for the listener ↔ speaker gain over the baseline runs from -0.06 to $+0.09$ and includes zero (one-sided $p \approx 0.26$), and every other family's interval includes zero as well. The strength of the claim is the consistency of the pattern across metrics and configurations, not any single delta. Impurity importance is biased toward the high-dimensional CLIP block, about 0.63 of the total, simply because it contributes the most columns, and permutation importance is near zero for every feature at $n = 89$, so no single feature is reliably pivotal. Among the interpretable structural families, listener head-motion dynamics, yaw velocity and its variability, carry the most weight, consistent with orientation and attention being the readable part of a silent listener's engagement. We therefore read the ablation, not the importance scores, as the evidence for what helps.

4.3 Speaker Prosody Partly Predicts Listener Engagement

The speaker side of the interaction also carries information about the listener's state, concentrated narrowly rather than spread across measures. Of the many prosodic measures we extract, almost none help on their own: as a standalone family, speaker prosody is a weak predictor (macro-F_1 0.306), and voice quality and turn-taking context add nothing on top of pitch. Within pitch, nearly all of the engagement-relevant signal resides in a single measure, the *terminal slope*: the rate at which the speaker's pitch rises or falls over the final half-second of voiced speech. We measure pitch in semitones relative to the window's median,

which makes it comparable across speakers (§3.2), so the terminal slope is a value in semitones per second (st/s). Averaged within each engagement class, this slope rises for windows where the listener is *Engaged* ($+1.5$ st/s) and *Intermittent* ($+2.6$ st/s), but is flat or falling where the listener is *Disengaged* (-3.6 st/s): engaged listeners co-occur with rising speaker intonation, disengaged ones with falling intonation. Because the transcripts carry no punctuation, this question-like intonation is recoverable only from the acoustics. We read it as a speaker-side correlate that partly predicts the listener's state: rising intonation is characteristic of questions, which typically invite a reaction, so a listener being addressed with a question is more likely to respond and appear engaged. Because the speaker addresses the group rather than the specific listener (§3.2), the cue indexes local interactional context rather than a signal aimed at the listener.

4.4 Label Scarcity Makes the Task Effectively Binary

The three-class results raise a prior question: is engagement simply hard to read from behaviour, or is the three-class *labelling* the obstacle? The middle *Intermittent* class is the prime suspect: it is where annotators disagree the most (§3.1), and the unsupervised exploration recovers only two natural groups (§4.1). If that ambiguous class is what holds performance down, merging it away should make the task markedly easier; if the features are simply too weak, no regrouping will improve the predictions. We therefore collapse the three labels into a binary contrast in three different ways, re-run the classification experiments, and compare each scheme against its own majority-class macro-F_1 floor (Table 2). The binary reductions differ only in their treatment of the Intermittent class. *Engaged vs. Non-active* groups Intermittent with Disengaged, while *Active vs. Disengaged* groups Intermittent with Engaged to form the *Active* class. *Engaged vs. Disengaged* excludes Intermittent from the evaluation and retains only high-confidence samples labelled as Engaged or Disengaged (E | D; $n = 61$).

The finding is clean: engagement is learnt more effectively as a binary *Engaged vs. Non-active* distinction. *Engaged vs. Non-active* reaches macro-F_1 0.634, $+0.30$ over its floor, the only wide, stable margin in the table. Every grouping that instead hinges on the *Disengaged* class collapses toward chance, whether the middle class is merged with Engaged (*Active vs. Disengaged*, 0.500 vs 0.447) or dropped (*Engaged vs. Disengaged*, 0.450 vs 0.419); the full three-class task clears its floor only modestly (0.378 vs 0.220). The paired bootstrap agrees: only *Engaged vs. Non-active* has a margin over its floor whose 95% interval excludes zero ($[+0.02, +0.28]$, one-sided $p = 0.015$), whereas the intervals for the *Disengaged*-dependent schemes reach down to zero.

Label scheme	Grouping	Macro-F_1	Floor	Acc.
Three-class	E | I | D	0.378	0.220	0.505
Engaged vs. Non-act.	E | I + D	**0.634**	0.336	0.641
Active vs. Diseng.	E + I | D	0.500	0.447	0.820
Engaged vs. Diseng.	E | D	0.450	0.419	0.722

Table 2: Label-scheme sensitivity (listener-only + speaker-alignment features). Base classes: E = Engaged (44), I = Intermittent (28), D = Disengaged (17). *Grouping*: "|" splits different groups. *Floor* is the majority-class macro-F_1 baseline. Only *Engaged vs. Non-active* clears its floor by a wide margin; schemes requiring the *Disengaged* class sit at their floor.

Three readings converge: the construct (*Intermittent* was defined as the label for windows an annotator could not confidently call Engaged or Disengaged, §3.1), the unsupervised $k = 2$ structure (§4.1), and learnability (only the *Engaged vs. Non-active* cut yields a large, stable gain). The task is effectively binary here, bounded by the scarcity of confident *Disengaged* labels (17 of 89), not by the feature set.

5 Conclusion

We measured the engagement of silent meeting participants through behavioral *alignment*, the relation between the listener and the people they attend to, rather than the listener alone. On the GEHM corpus under leakage-free cross-validation, the listener $\leftrightarrow$ speaker relation gave the largest improvement over a listener-only model, in the direction predicted by the literature. The margin is within fold variance, so it is a directional result rather than an established effect, and the design cannot attribute it to coordination rather than to shared exposure to the same speaker. The firmer findings are structural: with few and ambiguous labels the three-class task is effectively binary, bounded by the scarcity of confident disengaged labels (17 of 89) rather than by the feature set, and the speaker's terminal pitch slope remains an interpretable correlate of listener engagement.

Limitations

The clearest limitation is statistical: our supervised analysis rests on only 89 labelled windows, and every difference we report is comparable to or smaller than how much macro-F_1 already varies across cross-validation folds simply from how the data is split (roughly 0.09; §4), so no single result in Table 1 is statistically significant on its own. A paired bootstrap over the held-out predictions (§3.3) puts the 95% interval for the listener ↔ speaker gain over the baseline at $[-0.06, +0.09]$, including zero. The listener ↔ speaker alignment headline is a hypothesis supported by a consistent pattern, not an established effect size; replication on a larger labelled set is the obvious next step.

A second limitation is what the design can say about alignment at all. The alignment features are correlational and computed within a single five-second window, between two people participating in the same meeting. We therefore cannot separate the listener's individual behaviour from their coordination with others in a causal sense. A high listener ↔ speaker similarity is consistent with three accounts our features cannot distinguish: the listener is attending to and coordinating with the speaker; both are reacting independently to a third participant or a shared event; or the speaker's behaviour drives the listener's. Since alignment is measured from synchronic behaviours, we also cannot recover the lead-lag asymmetry that would begin to separate these accounts and that much of the synchrony literature relies on (Delaherche et al., 2012). Ablation gains should be read as incremental predictive value under a specific feature combination, not as evidence that coordination itself contributes engagement signal over individual behaviour. Separating the accounts would need lagged or directional measures, and ideally temporally resolved evidence of the kind used in dyadic work on shared attention (Wohltjen and Wheatley, 2021).

The labels are a third constraint: four non-expert annotators reached only modest agreement (Fleiss' $\kappa = 0.42$), mostly on the middle *Intermittent* class, which was defined as a region of annotator uncertainty and behaves as one (§3.1, §4.4). The binding scarcity, however, is the 17 confident *Disengaged* windows; with so few examples we cannot tell whether disengagement is genuinely hard to read from behaviour or simply under-sampled.

The listener ↔ speaker relation, our central result, is also measured under partial coverage: the active speaker is identified for only about 64% of windows, with the rest imputed to the training-fold median (§3.3). This keeps the protocol leakage-free but dilutes the measured contribution of the alignment and prosody families; a more robust speaker attribution, for instance from diarised audio, might sharpen the signal we observe.

Every feature family is also computed strictly within the target window, with no representation of the listener's own behaviour in earlier windows beyond the small amount of turn-taking history already in the speaker features (§3.2). Because engagement plausibly persists or drifts across turns rather than resetting every five seconds, a short history of the listener's own recent behaviour might recover signal that a single window cannot.

Finally, the findings are tied to one setting: GEHM consists of remote, gallery-view Zoom meetings in English, where the mutual-gaze and proxemic cues of co-present interaction are absent or distorted, so the alignment and prosody patterns we report may not transfer to in-person meetings, other languages, or different recording conditions. We also treat the speaker's prosody as concurrent context rather than a signal directed at the listener, since the speaker generally addresses the group; the terminal pitch slope is therefore an interpretable correlate (§4), not a validated causal cue.

References

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

Frank J. Bernieri, J. Steven Reznick, and Robert Rosenthal. 1991. The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of Personality and Social Psychology*, 60(2):243–253.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365.

M. Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: A review. *Smart Learning Environments*, 6(1):1–20.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Maria Frank, Ghassem Tofighi, Haisong Gu, and Renate Fruchter. 2016. Engagement detection in meetings. *Preprint*, arXiv:1608.08711.

Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1):59–109.

Daniel Gatica-Perez, Iain McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest-level in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 489–492. IEEE.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. DAiSEE: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885. *Preprint*, arXiv:1609.01885.

Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Transactions on Signal and Information Processing*, 7:1–16.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. 2011. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117, Portland, Oregon, USA. Association for Computational Linguistics.

Rivka Levitan, Agustín Gravano, Laura Willson, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada. Association for Computational Linguistics.

Jia Li, Yangchen Yu, Yin Chen, Yu Zhang, Peng Jia, Yunbo Xu, Ziqiang Li, Meng Wang, and Richang Hong. 2024. DAT: Dialogue-aware transformer with modality-group fusion for human engagement estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2024. MultiMediate'24: Multi-domain engagement estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, pages 11377–11382. Association for Computing Machinery.

Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18)*, pages 153–164. Association for Computing Machinery.

Catharine Oertel, Cécile De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. 2011. Towards the automatic detection of involvement in conversation. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues: COST 2102 International Conference*, pages 163–170. Springer.

Catharine Oertel, Kenneth A. Funes Mora, Joakim Gustafson, and Jean-Marc Odobez. 2015. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*, pages 107–114, New York, NY, USA. Association for Computing Machinery.

Shay Ohayon and Ilanit Gordon. 2025. Multimodal interpersonal synchrony: Systematic review and meta-analysis. *Behavioural Brain Research*, 480:115369.

Patrizia Paggio, Manex Agirrezabal, Giulia Di Cristina, Bart Jongejan, and Costanza Navarretta. 2026. Multimodal entrainment and feedback in online group

meetings. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9139–9149, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Patrizia Paggio, Manex Agirrezabal, Costanza Navarretta, and Leo Vitasovic. 2024. Multimodal behaviour in an online environment: The GEHM zoom corpus collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11890–11900, Torino, Italia. ELRA and ICCL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Armand Pellet-Rostaing, Ronan Bertrand, Adrien Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science*, 5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Monisha Singh, Ximi Hoque, Donghuo Zeng, Yanan Wang, Kazushi Ikeda, and Abhinav Dhall. 2023. Do i have your attention: A large scale engagement prediction dataset and baselines. In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023*, page 174–182. Association for Computing Machinery.

Sophie Wohltjen and Thalia Wheatley. 2021. Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences*, 118(37):e2106645118.

Including Non-Manual Markers in Studies of American Sign Language Dialogue

Jillian "Jam" Stebbins[1], **Amber Hamilton**[2], **Emily Nover**[1], **Stephanie Patterson**[1], **Allison Fitch**[1], **Patrick Boudreault**[3], **Raja Kushalnagar**[3], **Cecilia O. Alm**[1]
[1]Rochester Institute of Technology
[2] Rochester Institute of Technology/National Technical Institute for the Deaf
[3]Gallaudet University
js2680@rit.edu, aih5790@rit.edu, en7367@rit.edu, slp9101@rit.edu, ahfgsh@rit.edu, patrick.boudreault@gallaudet.edu, raja.kushalnagar@gallaudet.edu, coagla@rit.edu

Abstract

Language technology resources are limited for signed languages. In order to improve these technologies, large amounts of labeled language data stored in annotated corpora are necessary. Many existing annotated corpora in American Sign Language (ASL) do not include non-manual markers (NMMs), i.e., language features that do not involve the hands. This work underscores the importance of including NMMs in sign language research by describing their plurifunctional nature in signed conversation, specifying their link to sign language prosody, examining the current state of ASL corpus research, and recommending a systematized method of annotation which captures the critical but complex variations in NMM production in ASL.

1 Introduction

American Sign Language (ASL) is a natural, visual language with articulatory and perceptual features distinct from those of spoken languages (Valli et al., 2011). While signed languages primarily utilize the same neurological and linguistic systems as other natural languages (Newman et al., 2015), the visual modality impacts many aspects of language production and comprehension and allows a unique opportunity for cross-linguistic comparison. At the same time, it introduces distinct considerations for the development of language processing technologies. For the roughly 500,000 deaf and hard-of-hearing ASL users in the United States (Mitchell and Young, 2022), technologies such as Automatic Sign Recognition and Text-to-Sign lag behind Automatic Speech Recognition and Text-to-Speech (Bragg et al., 2019).

Improving language technology for ASL and other signed languages requires specialized methodologies designed for the visual modality, as well as engagement with the Deaf community by understanding technology users' actual needs and supporting Deaf and hard-of-hearing researchers in leading roles (Boudreault et al., 2024). ASL technologies developed by hearing non-signers typically have cultural and linguistic blind spots, leading to calls from the Deaf community to not only collaborate with, but center the perspectives of Deaf and hard-of-hearing researchers (Desai et al., 2024; Bragg et al., 2019). In addition, the development of sign language processing technologies often requires different techniques than those used for speech processing; these methodologies may be as straightforward as utilizing video rather than audio to capture language samples, or as complex as adopting alternate models for utterance segmentation (Moryossef et al., 2023). In this opinion paper, we bring attention to one set of ASL linguistic features that have been critically understudied in the computational linguistics literature: non-manual markers (NMMs).

1.1 Non-Manual Markers (NMMs)

The articulators most often studied in signed languages are the hands (Hochgesang, 2022). Manual features such as the shape, movement, and orientation of a signers' hands contribute heavily to word identification and grammar, and are typically seen as producing the lexico-morphological signs. However, non-manuals, or language features that do not utilize a signer's hands, are also critical for producing and comprehending signed language (Crasborn and Van der Kooij, 2013). These features, including head, face, and body movement, may influence an utterance's lexical content or provide suprasegmental cues or contextual information.

NMMs in signed languages are plurifunctional (Bauer et al., 2026), meaning any given feature may serve multiple purposes. These functions may be structural or affective. Affective functions of non-manuals include some facial expressions which are used in both spoken and signed languages (Benitez-Quiroz et al., 2016). However, in signed languages,

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 89–95
October 29, 2026 ©2026 Association for Computational Linguistics

NMMs also serve as language-specific features that can be used to disambiguate morphemes, convey syntax, and structure information. For example, raised eyebrows in ASL may demonstrate surprise or shock, but are more commonly used to determine phrase boundaries, indicate topicalization or question type, or function as an emphatic prosodic cue (Gonzalez, 2011). One NMM may be used for multiple functions on a very short time scale, often within one utterance (Oomen and Roelofsen, 2023). Due to their suprasegmental functions, NMMs are often associated with prosody in sign (Sandler, 2010). For instance, a signer may use eyebrow movement to indicate intonational phrases (Dachkovsky and Sandler, 2009), or use body shift to convey focus or emphasis (Wilbur, 2020).

While NMMs are critical for producing and understanding ASL in dialogic contexts, they are difficult to incorporate into language technologies like computer vision for sign detection or even DeepMind's new sign-language-to-text (SL2T) technology (DeepMind, 2026). If the same feature, for instance, body tilt, is used both to demonstrate contrast between options and to indicate that a signer is relaying information from another party (role shifting), the ability of tools like OpenFace (Baltrusaitis et al., 2018) or MediaPipe (Lugaresi et al., 2019) to detect when that feature is present does not allow for a full depiction of how body tilt functions in the utterance, or what the meaning of that utterance is. Research involving sign language and the development of sign language technologies requires more nuance than these tools currently provide. For further examples, computational models might not be able to fully parse an ASL question if eyebrow movement is not included, or distinguish the sign NOT-YET from LATE if mouth and head movement are ignored. The limitations of existing technologies in registering and reproducing NMMs in signed languages (Zhang et al., 2025) creates a resource gap in technologies such as sign language recognition (SLR), sign language generation (SLG), and affective computing for signers (Bragg et al., 2019).

2 Defining Prosody

2.1 Multimodal Prosody

In general, prosody relates to the suprasegmental features of language, which are patterns that help with producing and understanding dialogic context. Prosody can also be described as conveying informational structure (Zora et al., 2025); that is, prosodic cues do not necessarily impact morphology or syntax but nevertheless change the meaning of an utterance. Analyzing prosody typically involves examining duration, energy, and intonation (Haralambous, 2024), and all of these aspects are critical to fully relay any conversational message.

In spoken languages, prosodic cues appear primarily through features such as pitch, volume, speed, and pauses (Sandler, 2010). To indicate a yes or no question (i.e., to structure information) in English, for example, a speaker usually raises their pitch at the end of an utterance. However, prosody itself is not tied to any specific modality. Prosodic cues appear in all languages (and some nonlinguistic communication systems, as discussed in Applebaum et al., 2014) to allow dialogue partners to interpret what is actually meant by a statement. This is the case whether the prosodic cue or the language itself is spoken, signed, or tactile.

While most research on prosody is conducted on sound-based prosodic phenomena such as speech focus (Zora et al., 2025), prosody is not strictly tied to sound even in spoken languages (Prieto et al., 2024). Behaviors such as nodding (Frid et al., 2019) or holding the palms up (Lepeut and Shaw, 2025) may have some prosodic function in spoken languages, indicating focus, pausing, or other suprasegmental information. For visual languages like ASL, these same features are plurifunctional, so may be lexico-morphemic, prosodic, or affective depending on the utterance and its context.

2.2 Prosody in Sign

Prosodic cues are critical in dialogic interactions, and can indicate communicative intent, emotional or affective state, utterance boundaries, and other global contextual information. In ASL, energy and intonation are expressed not by changes in loudness and pitch, but through other articulators (Dachkovsky and Sandler, 2009). For example, frustration or anger (i.e., high energy) may be expressed by making a signed utterance larger. Many of these prosodic cues are non-manual: eyebrow, mouth, head, and body movement are all linked to prosody in signed languages (Crasborn and Van der Kooij, 2013).

Because signed languages are primarily in the visual modality, non-manual articulators serve more purposes than the same cues in spoken languages. Due to this previously discussed plurifunctionality, some aspects of ASL prosody may be difficult to

distinguish from syntax or even nonlinguistic gesture (Kita and Emmorey, 2023). Features identified as prosodic cues, such as eyebrow raises, often also serve other functions (i.e., conveying grammatical structure or affect) depending on the context of the utterance. This is a complexity that language processing technologies are currently less equipped to process (Orr et al., 2024).

However, there is evidence that first-language signers are able to accurately parse intonational phrases in ASL using as few as three prosodic cues (Gonzalez, 2011). They are also sensitive to a more selective array of cues (in the prior study, pauses, holds, and eyebrows) when parsing ASL clauses than signers of other languages, second-language signers, and non-signers, who all utilized a wider variety of prosodic cues to make the same determinations (Gonzalez, 2011). Other research has shown that, when signing speed is increased, certain NMMs change more drastically than others in both number of appearances and duration (Wilbur, 2009). Therefore, it is likely that native signers produce and attend to specific patterns of prosodic cues based on the function of an utterance. If language technologies gain the ability to identify and analyze the same cues as signers, improvements will be made in recognizing, processing, and producing ASL utterances.

3 Annotated Corpora in ASL

3.1 Existing Annotation Tools and Corpora

For computational models to learn how language is used in different contexts (e.g., by different speakers and in different locations) and to represent these differences accurately, especially for languages with fewer speakers or resources, the annotated corpora that provide the data needed for these models should be high-quality and reliable (Jurafsky and Martin, 2026). Annotated ASL corpora with clear structure can become benchmarks that train and evaluate novel systems, in addition to contributing to model improvements for tools trained on other ASL resources such as SL2T (DeepMind, 2026). For NMM annotations, human annotators should make consistent judgments about subtle face, head, and body movements that may be difficult to distinguish from one another.

Applications like OpenFace (Baltrusaitis et al., 2018), MediaPipe (Lugaresi et al., 2019), and OpenPose (Cao et al., 2021) have more general computer vision use than sign language analysis. Although these tools can demonstrate limited utility for automatically analyzing NMMs (Kuznetsova et al., 2021), they are not designed specifically for sign language applications and are missing key features that are necessary for processing sign language, such as the needed movement tracking for the tongue. This highlights the necessity of continuing to engage human annotators in the process of developing an improved annotation scheme. Existing automated tools, many of which utilize Action Units (AUs) from the Facial Action Coding System (Ekman and Friesen, 1978) to describe the motion of muscles in the face, do not necessarily help the annotation process for multiple reasons. First, facial AUs are designed and often used for emotion recognition within affective computing (Baltrusaitis et al., 2018). While there is some overlap between affective and structural non-manuals in signed languages (Benitez-Quiroz et al., 2016), it is not complete. Facial AUs may not capture the nuances of syntactical and prosodic NMMs. For example, ASL utilizes several mouth and tongue morphemes which a human annotator can distinguish, but current facial AU annotation schemata or automatic pose estimation tools lack such coverage (Zadeh et al., 2016). In addition, these schemata and software are primarily designed for and trained on non-signers, meaning that they may harbor inconsistencies and likely do not capture cultural differences in signers' non-manual production, including for affective expression. Accordingly, a standardized annotation system is important for ensuring consistency for both human annotators and future automatic annotation models, and for enabling cross-corpus comparisons and reliable model training.

Despite rapid improvements in the past few years, current annotations of ASL corpus data still focus on manual over non-manual cues. Many citation-form corpora, which have isolated dictionary-style tokens of individual signs, do not include contextual non-manual information. For example, ASL SignBank is an annotation tool meant to be used by fluent ASL researchers to make videos in ASL machine-readable (Hochgesang et al., 2019). It has over 3,000 sign-gloss pairs and is the gold-standard for a citation-form ASL corpus. Videos in SignBank are intentionally "unmarked" (Hochgesang, 2022), meaning that there are no NMMs involved, in order to maintain the dataset as a baseline for individual manual signs. While this is beneficial for standardization, it is

not sufficient to use corpora like this for developing language processing technologies that capture real-world NMMs use. The team behind SignBank has cautioned against using it for machine learning-based development due to this and other reasons (Hochgesang, 2026).

Continuous-form corpora, on the other hand, cover naturalistic signing in full and often include NMMs, even if they are not yet annotated as such. These corpora are used for different purposes than citation-form corpora and can include more contextual information. For example, one of the richest datasets is How2Sign, which includes multiple modalities, such as gloss, pose, and depth information (Duarte et al., 2021). However, How2Sign currently lacks NMM annotations. Other datasets serve specific purposes such as Sign Language Recognition (SLR) or Sign Language Translation (SLT), but there are no standardized annotation conventions for NMMs between these corpora. The current systematized conventions for annotating ASL, SLAASh (Hochgesang, 2022), focus predominantly on manual features, without many distinctions for NMMs (and specifically excludes from analysis ASL mouthings such as MM or PUFF-CHEEK) (Hochgesang and Lillo-Martin, 2020). The inclusion of more detailed NMM annotations into existing or novel ASL corpora will allow these datasets to become more complete and nuanced representations of signed languages.

3.2 Importance of Including NMMs in Annotated ASL Corpora

Because NMMs serve so many functions in signed languages, including as prosodic cues, omitting them from a conversation may obfuscate its meaning at any level of analysis. At the utterance level, for instance, signers consistently use NMMs to indicate permission, an explanation, a command, or advice (Brentari et al., 2018). However, not all ASL prosodic cues are non-manual and while NMMs serve many prosodic functions, they also have syntactic, morphological, and affective functions.

Some of the variation in NMM use is mostly standard across signers; for example, the phonological difference between NOT-YET and LATE depends on mouth and head movement, and most signers will sign the difference the same way. PUFF-CHEEK, for another example, is a mouth morpheme that may serve an adjectival function, impacting the lexical content of an utterance. NMMs that serve prosodic or affective functions are less standardized, and the same utterance might vary with NMM use from one signer to another (Valli et al., 2011), or from the same signer between contexts (Gonzalez, 2011). The variability of prosodic NMMs makes them more complex to annotate, but properly understanding them is a critical component of sign language comprehension. This complexity underscores the necessity of expanding NMM annotations when developing corpora used for training computational models of ASL.

There is still much to learn about NMMs, particularly their role in prosody, and current language technologies for ASL would benefit from novel and expanded corpus resources. For instance, automatic annotation systems for sign language struggle to catch specific features of NMMs, such as nose folding, forehead wrinkles, and tongue/teeth use (Hrúz et al., 2011). Annotation is further complicated whenever a signer's hands block NMMs from being visible on the face (Hrúz et al., 2011). In addition, NMMs add semantic richness to sign languages (Quesada et al., 2016; Orr et al., 2024), and existing facial action unit and facial gesture recognition tools built for emotion identification do not perform adequately on NMM-relevant tasks without additional human effort (Kuznetsova et al., 2021). Capturing NMMs reliably will require both more data and ASL-native human annotators; corpora developed this way can then support further ASL prosody research and NMM-aware models.

4 Discussion and Recommendations

The shortage of high-quality, annotated ASL data at scale remains a challenge for the development of accurate and relevant ASL processing technologies. Current corpus annotations are focused primarily on the manual articulators (Hochgesang, 2022), and current annotation tools are focused on features that are more relevant to affective computing than language processing applications. It is crucial to ensure annotation schemata are expanded to adequately capture the language-specific features of ASL, including NMMs.

An expanded annotation scheme for NMMs in ASL is necessary to model the behavior of different NMMs and their relationships to each other using quantitative representations. In addition, any computational models developed for the purpose of improving language technology will rely heavily on standardized annotations of large corpora. We note the importance of the creation of a common,

detailed template to annotate NMMs, especially those with links to ASL prosody. This scheme can be utilized to annotate existing or novel continuous-form corpora and improve the ability of language technologies to parse NMMs in ASL.

We argue that there should be additional nuance in these NMM annotations, such as including adjectival mouth morphemes, syntactic and affective eyebrow raises, and prosodic head and body tilt. Detailed annotations for NMMs will improve the utility of sign language data, especially for naturalistic dialogue between two or more signers.

4.1 Future Work

This paper is part of an ongoing project which has begun to collect a corpus of spontaneously-produced continuous signing from fluent signers. The corpus will utilize an expanded annotation scheme focusing on NMMs, which will be useful for computational models that process ASL, such as automatic annotation of other corpora. Future work should examine the implications of more detailed NMM annotations for machine learning-based models that train on annotated sign language data. The ability to model relationships between NMMs could potentially improve the performance of sign language technologies. Further, future work requires additional research on the distinction between different functions of prosodic NMMs in ASL, as a more robust and detailed protocol for the documentation and differentiation of NMMs will allow for better analysis and the use of NLP techniques in parsing and predicting NMMs in ASL.

5 Conclusion

This opinion paper asserts the importance of including NMMs in research related to ASL, especially due to their links to sign prosody. We discuss the state of ASL corpus research and advocate for a standardized method of annotating NMMs in projects designed to develop and improve language technologies. Our goal is to encourage corpus resource creation that can better capture the language-specific features of ASL so that emerging sign language technologies will have improved accuracy, allowing ASL tools to better serve Deaf and hard-of-hearing users. To support the conclusions found, the ongoing project relies on a strong base of knowledge from the Deaf sign language research community. Future works proposed are also relevant to researchers across many fields from linguistics to psychology to computer science, and especially for those who work on ASR, TTS, SLR, SLG, and other language technologies.

6 Limitations

This set of recommendations, while solidly based in the ASL linguistics and sign technology literatures, is only intended to support the study of NMMs in ASL. Other signed languages have different language-specific features and, while spoken languages may also have visual prosody, the conclusions drawn above do not generalize to languages other than ASL or types of prosody that are conveyed through articulators other than NMMs (i.e., the hands or vocal tract). Despite the fact that human annotations of NMMs are critical for the continued development of language technologies, there is a limit to the features that can be reliably coded. Some NMMs, such as blinking, squinting, and nose wrinkling, are subtle and difficult for human annotators to identify consistently. In addition, the function of some NMMs may be very difficult to classify in any given utterance, even for highly fluent or native annotators. This is a further challenge to the development of a broadly standardized annotation methodology. Finally, while ASL technologies necessitate improvement, this does not necessarily mean that it is impossible to do so without NNM annotations.

7 Ethical Considerations

While this work is an opinion paper and does not include data from human subjects participants, ethical considerations remain. Any project discussing sign language is positioned against the backdrop of a history of suppression of signed languages and the current realities of language deprivation for many deaf and hard-of-hearing children (Hall et al., 2017). This harm is, in part, the result of hearing academics who have operated under ableist frameworks without input from Deaf individuals–those who actually use the language (Hill and Hall, 2025). To reduce threats of adding to this harm, the ongoing project is being conducted by a team of Deaf and hearing researchers who operate in a bilingual (ASL/English) research space, and who developed the project collaboratively. Data collection sessions for the ongoing corpus project always involve at least one Deaf researcher, and all researchers interfacing with participants are users of ASL.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Award No. 2429899 and 2429900. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Lauren Applebaum, Marie Coppola, and Susan Goldin-Meadow. 2014. Prosody in a communication system developed without a language model. *Sign Language & Linguistics*, 17(2):181–212.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, Xi'an. IEEE.

Anastasia Bauer, Sonja Gipper, Tobias-Alexander Herrmann, and Jana Hosemann. 2026. Rethinking linguistic feedback: A modality-agnostic and holistic approach to multimodal addressee signals in spoken and signed dyadic interaction. *Glossa: A Journal of General Linguistics*, 11(1):18539.

C. Fabian Benitez-Quiroz, Ronnie B. Wilbur, and Aleix M. Martinez. 2016. The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150:77–84.

Patrick Boudreault, Muhammad Abubakar, Andrew Duran, Bridget Lam, Zehui Liu, Christian Vogler, and Raja Kushalnagar. 2024. Closed Sign Language Interpreting: A Usability Study. In Klaus Miesenberger, Petr Peňáz, and Makoto Kobayashi, editors, *Computers Helping People with Special Needs*, volume 14751, pages 42–49. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, Pittsburgh PA USA. ACM.

Diane Brentari, Joshua Falk, Anastasia Giannakidou, Annika Herrmann, Elisabeth Volk, and Markus Steinbach. 2018. Production and Comprehension of Prosodic Markers in Sign Language Imperatives. *Frontiers in Psychology*, 9:770.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186.

Onno Crasborn and Els Van der Kooij. 2013. The phonology of focus in sign language of the netherlands1. *Journal of Linguistics*, 49(3):515–565.

Svetlana Dachkovsky and Wendy Sandler. 2009. Visual intonation in the prosody of a sign language. *Language and Speech*, 52(2-3):287–314. PMID: 19624033.

Google DeepMind. 2026. Putting sign language AI into users' hands. Section: Models.

Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 54–65, Torino, Italia. ELRA and ICCL.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Paul Ekman and Wallace V. Friesen. 1978. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*.

Johan Frid, Malin Svensson Lundmark, Gilbert Ambrazaitis, Susanne Schötz, and David House. 2019. Investigating visual prosody using articulography. *Digital Humanities in the Nordic and Baltic Countries Publications*, 2(1).

Carolina Gonzalez. 2011. *Perception of Prosody in American Sign Language*. Purdue University.

Wyatte C. Hall, Leonard L. Levin, and Melissa L. Anderson. 2017. Language deprivation syndrome: a possible neurodevelopmental disorder with sociocultural origins. *Social Psychiatry and Psychiatric Epidemiology*, 52(6):761–776.

Yannis Haralambous. 2024. *A Course in Natural Language Processing*. Springer International Publishing, Cham.

Corinna S. Hill and Wyatte C. Hall. 2025. Peeling back the layers: contextualising the oralist narratives of listening and spoken language in deaf education and early intervention. *Oxford Review of Education*, pages 1–18.

Julie Hochgesang. 2022. SLAASh ID Glossing Principles, ASL Signbank and Annotation Conventions, Version 3.2.

Julie Hochgesang, Onno Crasborn, and Diane Lillo-Martin. 2019. ASL Signbank.

Julie Hochgesang and Diane Lillo-Martin. 2020. Sign language acquisition, annotation, archiving and sharing (SLAAASh).

Julie A. Hochgesang. 2026. *ASL Signbank*, 1 edition. Cambridge University Press.

Marek Hrúz, Zdeněk Krňoul, Pavel Campr, and Luděk Müller. 2011. Towards Automatic Annotation of Sign Language Dictionary Corpora. In *Text, Speech and Dialogue*, pages 331–339, Berlin, Heidelberg. Springer.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

Sotaro Kita and Karen Emmorey. 2023. Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology*, 2.

Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman. 2021. Using computer vision to analyze non-manual marking of questions in KRSL. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 49–59, Virtual. Association for Machine Translation in the Americas.

Alysson Lepeut and Emily Shaw. 2025. Interactive Functions of Palm-Up: Cross-Linguistic and Cross-Modal Insights from ASL, American English, LSFB and Belgian French. *Languages*, 10(9):239.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A framework for building perception pipelines. *Preprint*, arXiv:1906.08172.

Ross E Mitchell and Travas A Young. 2022. How many people use sign language? a national health survey-based estimate. *The Journal of Deaf Studies and Deaf Education*, 28(1):1–6.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Aaron J. Newman, Ted Supalla, Nina Fernandez, Elissa L. Newport, and Daphne Bavelier. 2015. Neural systems supporting linguistic structure, linguistic experience, and symbolic communication in sign language and gesture. *Proceedings of the National Academy of Sciences*, 112(37):11684–11689.

Marloes Oomen and Floris Roelofsen. 2023. Biased polar question forms in Sign Language of the Netherlands (NGT): Two functions of headshake. *FEAST. Formal and Experimental Advances in Sign language Theory*, 5:156–168.

Hayden Orr, Michael Peechatt, and Cecilia Ovesdotter Alm. 2024. MULTICOLLAB-ASL: Towards affective computing for the deaf community. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, New York, NY, USA. Association for Computing Machinery.

Pilar Prieto, Núria Esteve-Gibert, and Stefanie Shattuck-Hufnagel. 2024. Towards a novel conceptualization of prosody that accounts for spoken and visual signals: The modality-neutral prosodic framework hypothesis. *Gesture*, 23(1-2):119–159.

Luis Quesada, Gabriela Marín, and Luis A. Guerrero. 2016. Sign language recognition model combining non-manual markers and handshapes. In *Ubiquitous Computing and Ambient Intelligence*, pages 400–405, Cham. Springer International Publishing.

Wendy Sandler. 2010. Prosody and syntax in sign languages. *Transactions of the Philological Society*, 108(3):298–328.

Clayton Valli, Ceil Lucas, Kristin J. Mulrooney, and Miako Villanueva. 2011. *Linguistics of American Sign Language, 5th Ed.: An Introduction*. Gallaudet University Press.

Ronnie B Wilbur. 2009. Effects of varying rate of signing on asl manual signs and nonmanual markers. *Language and Speech*, 52(2-3):245–285.

Ronnie B. Wilbur. 2020. Non-manual markers: Theoretical and experimental perspectives. In *The Routledge Handbook of Theoretical and Experimental Sign Language Research*. Routledge. Num Pages: 36.

Amir Zadeh, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Convolutional Experts Constrained Local Model for Facial Landmark Detection. *arXiv preprint*. Version Number: 5.

Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025. Towards AI-driven Sign Language Generation with Non-manual Markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26, Yokohama Japan. ACM.

Hatice Zora, Helena Bowin, Mattias Heldner, Tomas Riad, and Peter Hagoort. 2025. Lexical and Information Structure Functions of Prosody and Their Relevance for Spoken Communication: Evidence from Psychometric and Electroencephalographic Data. *Journal of Cognitive Neuroscience*, 37(10):1633–1665.

Do Motion Tokenizers for Co-Speech Gesture Generation Encode Gesture Semantics?

Varsha Suresh[1,2]*, Divij Jain [3]*†, Jia Liu[1], M. Hamza Mughal[2], Vera Demberg[1,2]

[1]Saarland University, [2]Max Planck Institute for Informatics, Saarland Informatics Campus, [3]IIT Delhi

vsuresh@mpi-inf.mpg.de, cs5240052@iitd.ac.in,

jili00008@stud.uni-saarland.de, mmughal@mpi-inf.mpg.de, vera@lst.uni-saarland.de

Abstract

Discrete motion tokenizers encode motion as atomic units and are widely used for co-speech gesture generation. It remains unclear which motion properties, especially those relevant to gesture semantics, are recoverable from these codebooks. We probe a reconstruction-trained codebook using 19 co-speech gesture descriptors spanning from raw motion to abstract communicative function. Results show that geometry and handedness are readily decodable from token embeddings, while motion category is only weakly decoded despite showing systematic differences in discrete code usage. This gap between reconstruction quality and descriptor decodability suggests that reconstruction objectives alone do not guarantee that gesture semantics are captured, and that evaluating codebooks on such properties can guide the design of more semantic motion tokenizers.

1 Introduction

Co-speech gesture generation increasingly relies on discrete motion codebooks, representing continuous motion as sequences of learned tokens (van den Oord et al., 2017; Zhang et al., 2023; Jiang et al., 2023; Liu et al., 2024). These codebooks are trained with reconstruction objectives that average error across all joints and frames, favoring motion that is natural and fluid rather than motion that is semantically distinctive. Fine-grained movement, such as finger configuration, contributes only marginally to this loss, since it is overshadowed by larger-magnitude body motion, even though it can be communicatively important (Brentari, 1998). Since higher-level properties like motion category and communicative intent are partly expressed through such fine-grained motion, it is unclear whether these atomic units of movement are preserved in the codebook, which is subsequently used for co-speech gesture generation.

Co-speech gesture spans multiple levels of abstraction, from raw motion coordinates to descriptors such as spatial form, handedness, handshape, and communicative function (McNeill, 1992; Kendon, 2004; Brentari, 1998). Many of these higher-level descriptors are essential for a gesture to be semantically meaningful (McNeill, 1992; Kendon, 2004), yet it is unclear whether a codebook trained purely for motion reconstruction preserves them. In this paper, we aim to examine which of these properties are present in a reconstruction-trained gesture codebook, and which are lost.

To study this, we use the semantic gesture dataset from Ram et al. (2025). Building on this dataset and prior gesture annotation frameworks (Kipp et al., 2007; Hensel et al., 2025), we define a panel of 19 descriptors spanning raw motion to abstract communicative function. Some descriptors, such as joint-level geometry, are computed automatically from skeleton coordinates, while others, including handshape, handedness, and communicative intent, are manually annotated to capture aspects of gesture that are not directly recoverable from raw coordinates. Following prior work on probing learned representations (Alain and Bengio, 2016; Belinkov, 2022; Hewitt and Liang, 2019; Pimentel et al., 2020), we use these descriptors to characterize which gesture properties are recoverable from the codebook at two complementary levels: we apply linear and shallow nonlinear probes to mean-pooled token embeddings, and we compare code-activation distributions across descriptor classes to examine whether properties are reflected in patterns of discrete code usage.

Geometry and handedness are strongly decodable from pooled token representations, whereas handshape, motion category, and communicative intent are only weakly decoded. Motion category,

*These authors contributed equally to this work.

†Work done during an internship at Max Planck Institute for Informatics.

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 96–104
October 29, 2026 ©2026 Association for Computational Linguistics

however, shows systematic differences in discrete code usage. Handshape remains weakly recoverable under both analyses, suggesting that fine finger configurations, such as a cupped or pinched hand, are largely lost at the tokenization stage. We observe the same pattern on a second, architecturally distinct codebook, EMAGE (Liu et al., 2024), indicating that these results are not specific to a single tokenizer.

Our contributions are: (1) we introduce a systematic probing framework spanning 19 descriptors, from raw motion to communicative function, to characterize what reconstruction-trained gesture codebooks encode; and (2) Our results suggest that fine-grained, semantically important motion, such as handshape, may not be adequately preserved in a reconstruction-trained gesture codebook, even when the codebook achieves strong overall reconstruction accuracy.

2 Related Work

2.1 Gesture representations and descriptors

Gesture research has long analyzed movement form alongside communicative function (Efron, 1941; McNeill, 1992; Kendon, 2004; Goldin-Meadow, 2003). Computational approaches utilize raw trajectories (Kucherenko et al., 2019) to derive kinematic descriptors (Ferstl et al., 2020), or rely on manual annotations of location, orientation, handedness, and handshape (Kipp et al., 2007; Hensel et al., 2025). These methodologies parallel sign language phonology, which examines location, handshape, and movement (Stokoe, 1960; Brentari, 1998). At a higher level, gestures are categorized into semantic types (e.g., iconic, metaphoric) (McNeill, 1992; Kendon, 2004; Goldin-Meadow, 2003), which computational models often treat as distinct targets (Kucherenko et al., 2022). Recent models employ discrete tokenizers via vector quantization for general motion (Zhang et al., 2023; Jiang et al., 2023) and co-speech gesture (Liu et al., 2024). They are typically trained purely for reconstruction (van den Oord et al., 2017). While higher-level semantics are sometimes introduced externally (Zhang et al., 2024; Mughal et al., 2025), reconstruction performance alone does not clarify which interpretable properties the learned tokens inherently encode.

2.2 Probing learned representations

Probing tests whether specific properties are recoverable from learned representations (Alain and Bengio, 2016; Belinkov and Glass, 2019). We use linear probes (Hewitt and Liang, 2019; Pimentel et al., 2020) complemented by code-activation analysis to evaluate interpretable gesture features (Ghaleb et al., 2024), differentiating between simple accessibility and discrete structural encoding.

3 Data and Method

3.1 Descriptor Panel

We analyze the 19 gesture descriptors provided with the SEMANTIX dataset introduced by Suresh et al. (2026). We use the subset of 778 semantic gesture clips from TED, which originates from the gesture regions identified by Ram et al. (2025). Each clip is associated with descriptors covering physical gesture form and communicative function. These include automatically computed descriptors derived from the 3D skeleton motion, as well as manually annotated descriptors such as handshape, motion category class, and communicative function. In this paper, we retain the descriptor definitions and labels from the original study and organize the 19 descriptors into six groups to support our group-level probing and code-usage analyses. Table 2 summarizes the panel; full per-descriptor class labels are provided in Appendix A.

3.2 Codebook Architecture and Training

We analyze a two-stream vector-quantized autoencoder (VQ-VAE) (van den Oord et al., 2017) over a 38-joint 3D upper-body skeleton, partitioned into *body* and *hand* streams. Models are trained purely to minimize a position reconstruction and VQ commitment loss on a combined BEAT and TED gesture corpus. Our primary probing analyzes the best-reconstructing configuration: latent dimension $L = 256$, codebook size $K = 2048$, and temporal downsampling $d = 4$. Full architectural, configuration sweep, and training details are provided in Appendix A.1.2.

3.3 Probing Setup

We probe the codebook at two complementary levels: sequence-level probing of pooled token representations and code-level analysis of discrete code-usage patterns.

		Probing F1		Code usage		Reconstruction
Descriptor group	**N**	**Linear↑**	**Non-Linear ↑**	**MI ↑**	**Sig.**	**r(F1, MPJPE) ↓**
Motion dynamics	2	$0.49_{\pm 0.07}$	$0.52_{\pm 0.08}$	0.35	1/2	-0.26
Spatial geometry	9	$\mathbf{0.77}_{\pm 0.01}$	$\mathbf{0.74}_{\pm 0.03}$	1.07	9/9	-0.69
Handedness	1	$0.73_{\pm 0.03}$	$0.66_{\pm 0.05}$	0.77	1/1	$\mathbf{-0.89}$
Annotated spatial form	4	$0.30_{\pm 0.01}$	$0.31_{\pm 0.02}$	1.06	3/4	-0.48
Handshape	1	$0.16_{\pm 0.03}$	$0.20_{\pm 0.07}$	1.50	0/1	-0.67
Motion pattern and function	2	$0.11_{\pm 0.02}$	$0.12_{\pm 0.01}$	**3.62***	1/2	-0.45

Table 1: Group-level descriptor analysis. Linear and MLP probes use mean-pooled token embeddings; F1 is reported as mean ± SD across descriptors in each group. MI measures information in discrete code activations, and Sig. reports the number of descriptors with significant JS segregation. More negative correlations indicate that lower MPJPE is associated with higher probe performance. *This group-level MI is driven primarily by motion category (MI = 3.8, q = 0.006)

Group	**Descriptor**
Motion dynamics	Gesture size, Kinematics
Spatial geometry	Left/Right vertical, horizontal, depth, Left/Right elbow bend, Hand reach
Handedness	Handedness
Annotated spatial form	Hand orientation, Spatial radial, Spatial height, Spatial distance
Handshape	Handshape
Motion pattern and function	Motion category, Comm. intent

Table 2: Descriptor panel used in the main analysis.

Linear and nonlinear probing. Each gesture is represented using either mean-pooled codebook embeddings or normalized one-hot code histograms. For each descriptor, we train multinomial logistic regression with balanced class weights and report 5-fold cross-validated macro-F1 against majority-class and random-assignment baselines. We additionally train a shallow MLP with one 128-unit ReLU layer and L_2 regularization on the pooled embeddings, using identical folds and class balancing. This tests whether weak linear performance reflects nonlinear encoding while limiting probe capacity (Hewitt and Liang, 2019; Pimentel et al., 2020).

Code-level activation analysis. Each clip is represented as a normalized code-frequency histogram. We measure class separation using the mean pairwise Jensen–Shannon divergence between class-average usage profiles (Lin, 1991). Significance is assessed using label-permutation tests with Benjamini–Hochberg correction across descriptors within each stream. As a complementary measure, we sum the mutual information between each active code's presence and the descriptor label (Yang and Pedersen, 1997). Further details are provided in Appendix A.2.1.

Reconstruction fidelity. We test whether descriptor accessibility tracks reconstruction quality by correlating macro-F1 with mean per-joint position error (MPJPE) across models varying latent dimension, codebook size, and temporal downsampling. Further details are provided in Appendix A.1.2.

4 Results

4.1 Gesture Descriptor Probing

Table 1 summarizes descriptor accessibility from pooled token representations, structure in discrete code usage, and its relationship with reconstruction fidelity.

Accessibility from pooled representations. Probe performance shows a clear gradient from motion-grounded properties to higher-level gesture descriptors. Spatial geometry and handedness are the most accessible groups, reaching macro-F1 scores of 0.768 and 0.734, respectively. Performance decreases for annotated spatial form (0.299), hand configuration (0.155), and motion pattern and function (0.113). The descriptor-level comparison in Figure 1 shows the same pattern: geometric properties such as left and right depth exceed 0.89 macro-F1, whereas handshape, motion category, and communicative intent remain close to their chance baselines. A shallow MLP does not change this ordering; motion pattern and function remains effectively unchanged ($0.113 \rightarrow 0.112$). Its weak linear accessibility is therefore unlikely to reflect shallow nonlinear encoding alone.

Structure in discrete code usage. Activation analysis asks whether descriptor classes use different subsets of codes. Higher MI indicates that code activations are more informative about the

descriptor, while $q < 0.05$ indicates reliable separation after multiple-testing correction. Spatial geometry is significant for all nine descriptors, and handedness is also reliably separated. Most notably, motion pattern and function has the highest group-level MI (3.62), driven by motion category (MI $= 3.8$, $q = 0.006$), despite its weak pooled probing performance. This suggests that motion category is encoded in *which* codes are used, but is obscured by pooling. This does not hold for all higher-level properties: communicative intent ($q = 0.06$) and handshape ($q = 0.70$) show no reliable overall separation. Thus, weak pooled probing reflects hidden code-level structure for motion category, but limited evidence for handshape and most communicative intents.

Relationship with reconstruction metrics Negative correlations indicate that lower reconstruction error is associated with higher probe performance. This relationship is strongest for handedness ($r = -0.88$) and spatial geometry ($r = -0.69$), suggesting that reconstruction quality mainly tracks motion-grounded properties.

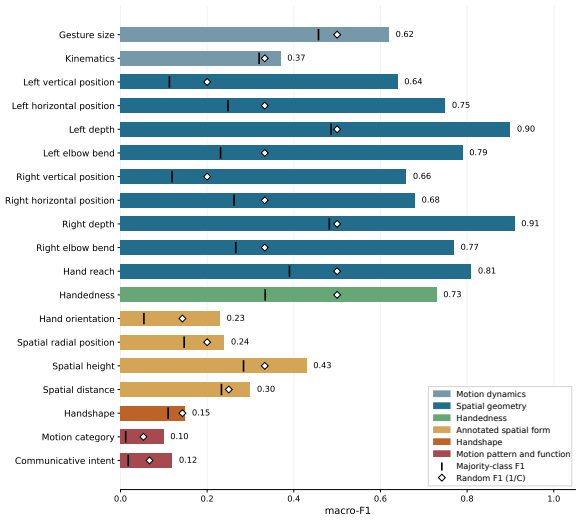


Figure 1: Probe macro-F1 for each descriptor

4.2 Generalization to Another Codebook

We repeat the analysis on EMAGE (Liu et al., 2024). As shown in Figure 2, both codebooks exhibit the same accessibility gradient: spatial geometry is strongest, while properties such as handshape remain weak. Although absolute scores differ slightly, the ordering of descriptor groups is stable across architectures. This consistency suggests that the accessibility gradient is not specific to a single tokenizer design.

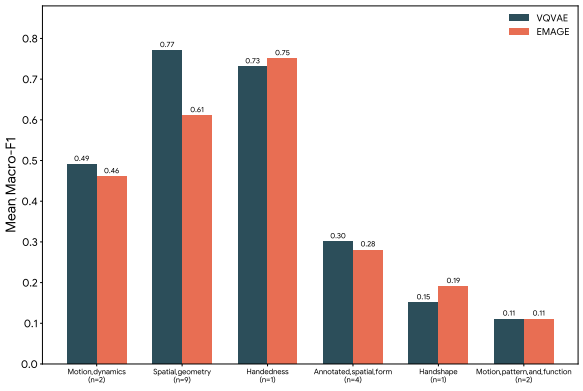


Figure 2: Probing macro-F1 by descriptor group, for EMAGE.

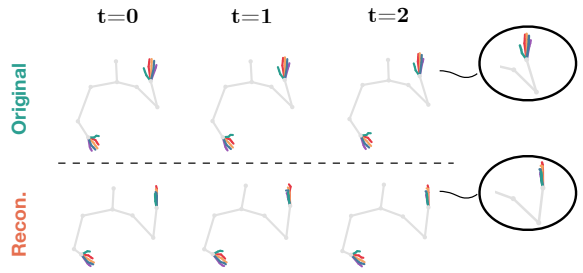


Figure 3: Qualitative comparison showing that reconstruction-only VQ-VAEs preserve upper-body geometry but average out fine-grained hand poses

5 Qualitative Case Study

Figure 3 illustrates a pattern that corroborates probing performance on communicative functions. VQ-VAE based tokenizers trained solely with a reconstruction loss, learn representations that decode into averaged-out motions. In particular, hands do not achieve same fidelity at reconstruction as compared to upper body joints, due to difference of scale between hands and upper body. Therefore, better performance on upper body results in strong accessibility of geometric descriptors. In contrast, poorly reconstructed hand poses depict different communicative functions, which eventually leads to subpar performance on gesture functions. These findings highlight that reconstruction losses do not incentivize learning of gesture semantics as part of the representation learning in codebooks.

6 Conclusion

Across two codebook architectures, geometric properties and handedness are readily recoverable, while handshape, motion category, and communicative intent remain weak. Better reconstruction and

finer temporal resolution mainly improve motion-grounded properties, suggesting that reconstruction quality alone is not a sufficient measure of semantic accessibility. Gesture codebooks should therefore be evaluated using both probing and code-level analyses, and future tokenizers should explicitly preserve communicatively relevant properties.

Limitations

Our analysis is limited to 778 semantic gestures from TED and two skeleton-based, reconstruction-trained tokenizers, so the observed accessibility gradient may not generalize to other domains, motion representations, or tokenizer objectives. In addition, we did not ablate the composition of the training corpus as the codebook was trained on BEAT and TED jointly, so the contribution of each dataset to descriptor accessibility remains untested. Manual descriptors, particularly handshape and communicative intent, are relatively sparse and may contain annotation subjectivity. Moreover, low probe performance indicates limited recoverability under the selected pooled representations and probe families; it does not establish that a property is entirely absent or inaccessible to a sequence-aware model. Future work should evaluate larger and more diverse datasets, use temporal probes, and test whether descriptor accessibility predicts the semantic quality of generated gestures.

Acknowledgments

This work is funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914."

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. MIT Press.

David Efron. 1941. *Gesture and Environment*. King’s Crown Press.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA)*.

Esam Ghaleb, Bulat Khaertdinov, Wim Pouw, Marlou Rasenberg, Judith Holler, Asli Ozyurek, and Raquel Fernández. 2024. Learning co-speech gesture representations in dialogue through contrastive learning: An intrinsic evaluation. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 274–283.

Susan Goldin-Meadow. 2003. *Hearing Gesture: How Our Hands Help Us Think*. Harvard University Press.

Laura B. Hensel, Stephanie Cheng, and Stacy Marsella. 2025. A richly annotated dataset of co-speech hand gestures across diverse speaker contexts. *Scientific Data*, 12:1748.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Michael Kipp, Michael Neff, and Irene Albrecht. 2007. An annotation scheme for conversational gestures: How to economically capture timing and form. *Language Resources and Evaluation*, 41(3):325–339.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA)*, pages 97–104.

Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2022. Multimodal analysis of the predictability of hand-gesture properties. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024.

Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1144–1154.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. 2025. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16578–16588.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Ashwin Ram, Varsha Suresh, Artin Saberpour Abadian, Vera Demberg, and Jürgen Steimle. 2025. Gesturecoach: Rehearsing for engaging talks with llm-driven gesture recommendations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the american deaf. Technical report, University of Buffalo.

Varsha Suresh, Mohammad Mahdi Abootorabi, Mohamed Salman, M Hamza Mughal, Christian Theobalt, Ashwin Ram, Jürgen Steimle, and Vera Demberg. 2026. Semantic motion anchors: Bridging motion and meaning in co-speech gestures. *arXiv preprint arXiv:2605.30608*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420. Morgan Kaufmann.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14730–14740.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)*, 43(4):1–17.

A Appendix

A.1 Additional Details: Data and Method

A.1.1 Descriptor Classes

Table 3 shows the detailed classes in each descriptor.

A.1.2 Codebook Architecture and Training

We analyse a two-stream vector-quantised autoencoder over a 38-joint 3D upper-body skeleton. The joints are partitioned into a *body* stream of 8 coarse joints and a *hand* stream of 30 finger joints. The two streams have separate encoders and codebooks and are quantised independently; a single shared decoder reconstructs the full pose. The hand joints are expressed relative to their wrist before encoding, meaning the hand stream encodes wrist-local finger configuration while the body stream encodes coarse-joint positions.

Each stream's encoder is a 1D temporal convolutional stack (kernel 4, stride 2) that downsamples time by a factor d. Each codebook is a vector-quantised bank trained with exponential-moving-average updates (decay 0.99) and dead-code reinitialisation. Tokenisation processes each clip in non-overlapping d-frame chunks, yielding one token per stream per chunk.

To map how representational granularity affects descriptor accessibility, we swept a $3 \times 3 \times 3$ grid of 27 models over $L \in \{64, 128, 256\}$, $K \in \{1024, 2048, 4096\}$, and $d \in \{4, 8, 16\}$. Every model is trained purely to reconstruct motion. The codebook is trained on a combined 3D gesture-motion corpus comprising the BEAT and TED datasets (processed at native frame rates of 30 FPS and 24–30 FPS respectively, balanced by duration). Models are optimised with the Adam optimiser (learning rate 1×10^{-3}, batch size 8) for up to 25 epochs (with early stopping), minimising a position reconstruction loss plus the VQ commitment loss weighted by $\beta = 1.0$.

A.1.3 Probing Setup

Table 4 reports descriptor-level linear probing results. The main pattern is consistent with the group-level results in the main paper. Automatic geometric descriptors and handedness are the strongest, while handshape, motion category, and communicative intent remain weak.

A notable trend across these descriptors is that the continuous codebook vector (Embedding) outperforms the code histogram (One-hot) for almost

Group	Descriptor	Annotation	C	Labels
Motion dynamics	Gesture size	Manual	2	normal; big
Motion dynamics	Kinematics	Manual	3	slow; normal; fast
Spatial geometry	Left vertical position	Automatic	5	below-waist; waist-level; torso-level; chest-level; above-head
Spatial geometry	Left horizontal position	Automatic	3	body-center; torso-side; extended-outward
Spatial geometry	Left depth	Automatic	2	at-torso; in-front-of-torso
Spatial geometry	Left elbow bend	Automatic	3	straight; slightly-bent; bent
Spatial geometry	Right vertical position	Automatic	5	below-waist; waist-level; torso-level; chest-level; above-head
Spatial geometry	Right horizontal position	Automatic	3	body-center; torso-side; extended-outward
Spatial geometry	Right depth	Automatic	2	at-torso; in-front-of-torso
Spatial geometry	Right elbow bend	Automatic	3	straight; slightly-bent; bent
Spatial geometry	Hand reach	Automatic	2	mid-reach; extended
Handedness	Handedness	Manual	2	one hand; both hands
Annotated spatial form	Hand orientation	Manual	7	inward; outward; toward center; up; down; side; forward
Annotated spatial form	Spatial radial position	Manual	5	inward; front; side; out; far out
Annotated spatial form	Spatial height	Manual	3	belt; chest; head
Annotated spatial form	Spatial distance	Manual	4	touch; close; normal; far
Handshape	Handshape	Manual	7	palm; cup; pinch; fist; bunch; fingers; index finger
Motion pattern and function	Motion category	Manual	19	arc-like trajectory; counting; diagonal; flick; glide; horizontal; long-tail; pointing; pull; push; rotate; shake; step; strike; sway; sweep; vertical down; vertical up; wave
Motion pattern and function	Communicative intent	Manual	15	contrast; discourse; emotion; emphasis; enumeration; listing; negation; other-reference; progression; quantification; relativity; self-reference; symbolise; temporal reference; uncertainty

Table 3: The 19-descriptor panel used for probing. Automatic descriptors are computed and discretized from skeleton coordinates, whereas manual descriptors are human annotated. C denotes the number of retained classes after removing labels represented by fewer than five clips.

all descriptors. The advantage is largest on graded spatial/positional descriptors (e.g., vertical reach bias, spatial distance), where a continuous vector successfully captures gradation that a discrete histogram naturally discards. One-hot representations only edge ahead on a handful of low-signal descriptors (like handshape), where the absolute F1 is already near the noise floor.

A.2 Additional Results

A.2.1 Code-Level Activation Procedure & Results

To rigorously evaluate whether a descriptor is reflected in the use of individual codebook entries, we employ an information-theoretic approach.

Jensen-Shannon divergence provides a symmetric and bounded metric for scoring how much the classes' profiles differ. The summed presence Mutual Information (MI) provides an additive measure of how much information the active vocabulary collectively carries about the label, avoiding the combinatorial infeasibility of estimating a full joint contingency table $I(Y;K)$ over all 2048 codes.

Table 5 reports the descriptor-level results. The data reveals that weak linear performance does not possess a single interpretation. Some descriptors leave strong code-level traces despite failing linear probing (e.g., motion category, $q = 0.006$), indicating the information lives in *which* discrete units are activated rather than their averaged vector geometry. Conversely, other abstract properties like communicative intent ($q = 0.063$) and handshape ($q = 0.697$) fail to reach broad code-level significance, indicating that these features are genuinely absent from the token representation.

A.2.2 Further examples for the code-level metrics.

To clarify the code-level metrics, we describe what they return in two extreme cases and in two real descriptors. In the first extreme, both classes of a binary descriptor use every code equally often. The two class-average profiles are then identical, so JS $= 0$, and because knowing which code fired tells us nothing about the class, MI $= 0$ as well. In the opposite extreme, each class uses only its own code and never the other's. The profiles share no codes, so JS reaches its maximum of 1 bit, and observing a single code reveals the class with certainty, so MI is maximal.

Real descriptors sit between these two extremes. For handedness, both classes share generic resting codes, but "one hand" clips concentrate on a few specific codes (e.g., hand code 746) while "both hands" clips use a separate set of bimanual codes. This partial overlap gives intermediate values (JS $= 0.31$, MI $= 0.766$), and the permutation test confirms the separation is far above what random labels would produce ($q = 0.006$). Left depth classes are heavily imbalanced (94% vs. 6%), so the minority profile is estimated from very few clips and looks spiky. JS weighs both profiles equally regardless of how many clips they come from, so it becomes inflated (JS $= 0.83$). MI instead accounts for how rare the minority class is, so it stays moderate (MI $= 0.493$). Reading JS together with

its permutation q-value and MI therefore prevents false discoveries caused by class imbalance.

Group	Descriptor	Cls.	Maj-F1	Rand-F1	Embed.	One-hot
Motion dynamics	Gesture size	2	0.457	0.500	**0.619**	0.562
Motion dynamics	Kinematics	3	0.320	0.333	**0.367**	0.340
Spatial geometry	Left vertical position	5	0.113	0.200	**0.639**	0.563
Spatial geometry	Left horizontal position	3	0.248	0.333	**0.749**	0.611
Spatial geometry	Left depth	2	0.486	0.500	**0.898**	0.884
Spatial geometry	Left elbow bend	3	0.231	0.333	**0.787**	0.669
Spatial geometry	Right vertical position	5	0.119	0.200	**0.663**	0.467
Spatial geometry	Right horizontal position	3	0.262	0.333	**0.679**	0.566
Spatial geometry	Right depth	2	0.482	0.500	**0.912**	0.858
Spatial geometry	Right elbow bend	3	0.266	0.333	**0.772**	0.487
Spatial geometry	Hand reach	2	0.390	0.500	**0.809**	0.751
Handedness	Handedness	2	0.334	0.500	**0.734**	0.664
Annotated spatial form	Hand orientation	7	0.054	0.143	**0.227**	0.201
Annotated spatial form	Spatial radial position	5	0.147	0.200	**0.236**	0.228
Annotated spatial form	Spatial height	3	0.284	0.333	0.434	**0.438**
Annotated spatial form	Spatial distance	4	0.233	0.250	**0.300**	0.297
Handshape	Handshape	7	0.110	0.143	0.155	**0.197**
Motion pattern and function	Motion category	19	0.012	0.053	**0.105**	0.083
Motion pattern and function	Communicative intent	15	0.018	0.067	**0.121**	0.101

Table 4: Descriptor-level linear probing macro-F1 with chance baselines. **Cls.** is the number of retained classes (≥ 5 samples). **Maj-F1** is the majority-class macro-F1 and **Rand-F1** is the random-assignment macro-F1 ($= 1/C$ in expectation). For near-balanced binary labels such as depth and handedness, Rand-F1 exceeds Maj-F1, so the higher of the two is the relevant chance level. The best of the two probe representations is shown in bold. Spatial geometry and handedness clear chance by large margins, while motion category, communicative intent, and handshape exceed chance only marginally.

Group	Descriptor	Cls.	Stream	MI	q
Motion dynamics	Gesture size	2	hand	0.397	**0.006**
Motion dynamics	Kinematics	3	hand	0.311	0.327
Spatial geometry	Left vertical position	5	hand	1.849	**0.006**
Spatial geometry	Left horizontal position	3	hand	1.056	**0.006**
Spatial geometry	Left depth	2	hand	0.493	**0.006**
Spatial geometry	Left elbow bend	3	hand	1.039	**0.006**
Spatial geometry	Right vertical position	5	hand	1.851	**0.006**
Spatial geometry	Right horizontal position	3	hand	1.024	**0.006**
Spatial geometry	Right depth	2	hand	0.575	**0.006**
Spatial geometry	Right elbow bend	3	hand	0.875	**0.006**
Spatial geometry	Hand reach	2	hand	0.865	**0.006**
Handedness	Handedness	2	hand	0.766	**0.006**
Annotated spatial form	Hand orientation	7	hand	1.802	**0.006**
Annotated spatial form	Spatial radial position	5	hand	1.191	**0.006**
Annotated spatial form	Spatial height	3	hand	0.534	0.208
Annotated spatial form	Spatial distance	4	hand	0.697	**0.021**
Handshape	Handshape	7	hand	1.499	0.697
Motion pattern and function	Motion category	19	hand	3.790	**0.006**
Motion pattern and function	Communicative intent	15	hand	3.450	0.063

Table 5: Descriptor-level code activation results using the stream with the highest mutual information for each descriptor. Bold q-values indicate FDR-corrected $q < 0.05$. Spatial geometry is consistently significant. Motion category is the main higher-level descriptor that shows a reliable code-level trace despite weak linear probing performance.

Looking for Objects and Looking for Help: Robot Gaze Elicits Human Cooperation

Anna Manaseryan
annamanaseryan@u.boisestate.edu
Computer Science
Boise State University

Thi Lan Anh Bui
anhbui@u.boisestate.edu
Computer Science
Boise State University

Casey Kennington
caseykennington@boisestate.edu
Computer Science
Boise State University

Abstract

Robots that search for objects on their own must decide what to do when the search fails, and how they signal that failure to a human partner can determine whether the interaction turns into collaboration. Prior work on human–robot interaction has studied nonverbal help-seeking and social gaze separately, but rarely compares gaze against a matched verbal failure signal inside a live, real-time embodied search task. We present an embodied dialogue system, built on the Retico incremental processing framework, where the Cozmo robot searches for a specified object and signals failure either verbally or through gaze towards the human partner. We first compare a rule-based planner against a vision-language model (VLM) planner that reasons over scene semantics, finding that the VLM planner succeeds far more often (87.2% vs. 65.9%). We then test whether gaze can close this gap without adding reasoning capability to the rule-based planner. Adding gaze as the stuck signal raises success to 87.87%, matching the VLM planner, while a verbal stuck signal only raises it slightly (68.75%). Results further show that this effect is largely implicit: participants helped more in the gaze condition without consistently reporting that they felt invited to do so. Our findings suggest that gaze can substitute for reasoning capability a robot lacks, helping it elicit assistance from a human partner.

1 Introduction

Multimodal, situated interaction is central to natural human–robot dialogue where people and robots coordinate not only through words, but also through gestures, emotion, and other information (McNeill and Kennington, 2020; Admoni and Scassellati, 2017; Novikova et al., 2015; Mutlu et al., 2012). In a setting where a robot explores a space for objects, the robot must signal a need for help to the human when the exploration fails, otherwise the human collaborator cannot know whether some kind of an intervention to help the robot will actually succeed. When a robot fails to find an object, the way it signals that failure matters. A verbal statement such as *I cannot find the object* communicates the need for help, but does not naturally open the floor for collaboration.

Prior work shows that nonverbal signals can act as requests for help during collaborative tasks (Cha and Matarić, 2016). One specific nonverbal signal — turning to face the person and holding the person's gaze — is a social act that can implicitly invite participation without requiring an explicit request (Mutlu et al., 2009; Cha and Matarić, 2016). *Eye gaze* is an important communicative modality that can invite participation from humans by shaping conversational footing (Mutlu et al., 2009), and other work shows that eye contact changes how later robot gaze is processed (Kompatsiari et al., 2018). It has also been shown that robot behaviors display emotion (Novikova et al., 2015; McNeill and Kennington, 2019) and that emotional expression encourages helping in collaborative tasks between robots and humans (McNeill and Kennington, 2020; Zhou and Tian, 2020). Taken together, these findings suggest that failure-triggered gaze toward a human partner coupled with an appropriate emotional behavior could elicit collaboration in a live object-search task. Following work on nonverbal help-seeking in human–robot collaboration (Cha and Matarić, 2016) and on social eye gaze in human–robot interaction (**HRI**) (Admoni and Scassellati, 2017; Mutlu et al., 2009), in this work, we task a human and a robot to collaboratively seek out objects, but when a robot is 'stuck' it uses eye gaze as a communicative act; i.e., at the moment the robot cannot find a target, a policy must select *how* to disclose failure, which is then realized as either speech or embodied orientation with gaze toward the human.

Our work uses a system built on the Retico incremental processing and 'robot-ready' spoken di-

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 105–116
October 29, 2026 ©2026 Association for Computational Linguistics

alogue system frameworks (Michael, 2020; Manaseryan et al., 2025). Our Experiment first establishes a baseline: Human participants tell the robot which object they want it to find and we use a language model to autonomously control a robot to explore objects in a confined space, which shows that a language model can generate plans better than a rule-based planner, but in both systems there are many cases where the robot is unable to find a specified object. That leads to our primary experiment where we ask: *How can the robot elicit help from the human when the robot cannot find the object?* In that experiment, two conditions share mid-search frustration displays; only the stuck signal differs: either a verbal 'give-up' ("can't find") or a physical turn toward the participant with held gaze. Thus emotion is held constant across conditions and cannot explain the between-condition difference; the contrast is the stuck cue itself. We find that the robot is able to complete more tasks (i.e., find more objects) when it is able to elicit help from the user both using speech and gaze, but more so using gaze, to the degree that the interactive help from human participants parities a more sophisticated planner used to seek objects.

Our contributions are as follows:

1. A real-time embodied dialogue system built on Retico (Michael, 2020; Manaseryan et al., 2025) that integrates automatic speech recognition, object detection and search planning, generative multimodal emotional behavior, and robot execution in a single end-to-end interactive loop.

2. A controlled comparison of visual planning strategies for physical object search task on Cozmo, establishing the live search setting on which Cozmo requests help when it cannot find the object.

3. A controlled comparison of gaze after failure versus a matched verbal give-up. We evaluate spontaneous collaborative guidance and how participants interpret the stuck cue as an invitation to help. We also observe whether directional hints after the stuck signal allow the robot to continue searches that would otherwise end.

In the following section we compare our work to others, then explain our methodology. We then explain two experiments where we (1) evaluate alternative planners for our object-search task, then (2) test whether gaze at task failure invites collaboration relative to a verbal baseline.. We then offer limitations and conclude.

2 Related Work

Social eye gaze is a communicative channel in HRI. Robots use gaze to manage attention, show engagement, and shape who is invited to take part (Admoni and Scassellati, 2017). Mutlu et al. (2009) showed that gaze cues alone can assign conversational footing (addressee, bystander, or overhearer), with participants following robot-assigned roles at high rates. Later work extended these ideas to turn-taking and dialogue structure with more than two people (Mutlu et al., 2012). Other evidence shows that looking at a person's face before a later gaze shift changes how that cue is processed (Kompatsiari et al., 2018), and that human gaze toward a robot often comes before joint action in collaborative assembly (Lavit Nicora et al., 2024). We build on this by treating the *robot's* gaze toward the human after failure as an invitation to collaborate once search has stalled.

A parallel line of work asks how robots should request help when collaboration breaks down. Cha and Matarić (2016) found that nonverbal help-request signals (light and sound at different urgency levels) change both helping behavior and how people see the collaboration for a non-humanoid robot. Other related work studies the reverse case. *Human* gaze can show that the person has noticed robot failure, including real-time detection from gaze patterns (Tabatabaei et al., 2025a,b). This work (specifically, when robot requests help when it cannot find the object) instead treats *robot-to-human* gaze itself as the failure signal.

Visual planning for physical robot tasks has moved from rule-based detection toward vision-language models that reason over scene semantics rather than geometry alone (Driess et al., 2023). Recent VLMs such as Qwen2.5-VL-7B (Bai et al., 2025) can perform object localization and spatial reasoning from raw scene context, without relying on bounding-box confidence thresholds. Experiment 1 compares this approach directly against a YOLO-based rule-based planner on a live object-search task.

Related to eye gaze, affect and emotion form a separate channel for eliciting help from humans. Emotionally expressive robots are more likely to receive collaboration in joint tasks, including help

from the human and task success time (McNeill and Kennington, 2020; Zhou and Tian, 2020). A specific emotion, *frustration* behaviors are also argued to matter for successful human–robot collaboration more broadly (Weidemann and Rußwinkel, 2021). In this work, we use emotional behaviors to improve the interactive qualities of our robot, though we do not compare an emotional vs. non-emotional robot display as the emotion (e.g., *frustration*). Emotional displays are a shared baseline that signals something is wrong, but not specifically that help is requested.



Figure 1: The Cozmo robot (Digital Dream Labs), the embodied platform used in both experiments.

3 Method

Robot Platform Both of our experiments use Cozmo, a small toy-like robot licensed by Digital Dream Labs (see Figure 1). Cozmo has a front-facing RGB camera (320×240 at 15 frames per second), an OLED screen for facial expressions, a movable head, a lift arm, and drive wheels for tabletop navigation. Control is made through the Cozmo Python SDK. These capabilities make it a suitable platform for multimodal face-to-face interaction and for our task: the camera supports vision-based search, the screen and body actuators support emotional behaviors, and the head and drive motors support physical orientation toward a human partner. It has also been shown in prior work that a non-verbal Cozmo elicited cooperation from humans based on how its beaviors displayed emotion (McNeill and Kennington, 2020), and it has also been shown that people view Cozmo as a child-level robot that is 2-3 years old in comparison with human cognition (Plane et al., 2018), which is important because children often use physical displays of emotion (e.g., frustration) and eye gaze as a signal of confusion and to elicit help from others before they are able to speak.

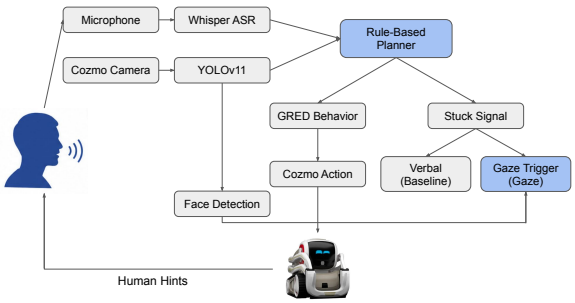


Figure 2: System architecture for our experiment. The rule-based planner receives both ASR and YOLO streams and coordinates all robot actions, including GRED emotional behaviors and the stuck signal (verbal 'stuck' state; gaze orientation in Gaze condition).

System Architecture Figure 2 shows the full system. Our system is built on the Retico incremental processing framework (Michael, 2020; Manaseryan et al., 2025), which connects modules by passing Incremental Units (IUs) as fine-grained units of information between modules (Schlangen and Skantze, 2011). Such fine-grained, incremental processing is an important property of natural dialogue between robots and people (Manaseryan et al., 2025). Two input modalities feed the central planner in parallel: (1) the speech modality, which uses a `Microphone Module` to stream raw audio to `Whisper ASR Module`, which emits word-level updates as speech arrives and a committed utterance once the participant finishes speaking, and (2) the vision modality, which uses a `Cozmo Camera Module` to read image frames from Cozmo's front camera and pass them to a *YOLOv11*[1] *Module*, which emits detected object classes, confidences, and bounding boxes. The planner receives both streams (transcribed speech and bounding box / object labels), maintains search state, and emits action commands to `Cozmo Action Module`, which serializes them in a queue and executes them one at a time through the `Cozmo Behaviors Module`. Actions include search movements, emotional behaviors, and spoken output. A separate `Face Detection Module` acts as another modality input that scans for the participant's face at startup and stores their location. In the experiment, this module is also used to enable Cozmo to angle its face towards and gaze at the human participant when needed.

[1] https://github.com/ultralytics/ultralytics

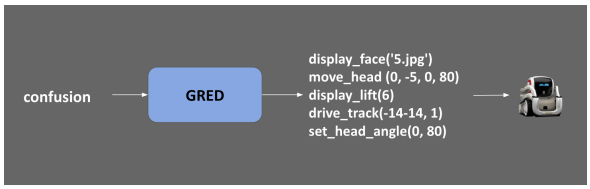


Figure 3: GRED maps an emotion label (here, confusion) to a sequence of Cozmo actions, including facial display, head motion, lift, and drive, which are then executed on the robot.

GRED Emotional robot behavior is produced by GRED (Generate Robot Emotional Displays) (Baral et al., 2025), a generative model built on GPT-2 Medium. GRED takes an emotion label and produces a novel sequence of parameterized Cozmo actions, including facial displays on an OLED screen, short vocalizations, driving patterns, head angles, and lift-arm movements, such that playing back those actions on the robot would make an observer think the robot is displaying the specified emotion. The model was fine-tuned on hand-designed behavior sequences across six grouped emotion categories, including Anger/Frustration and Joy/Hope. Each behavior is a sequence of symbolic action tokens (for example, `display_oled_face_image-face2 turn_in_place 30 230`). An example of input and output can be seen in Figure 3. At inference time, our search Planner module emits an emotion label to GRED and executes the returned behavior on Cozmo. In this work, the Planner Module uses `anger_frustration` during mid-search escalation when the target is still missing, and `joy_hope` when the object is found. In our experiments, both conditions share these GRED displays. Displaying emotions is very important as the robot needs to have a basic ability to display emotional behaviors or the evaluation would not be convincing as the robot would be perceived by humans as "too robotic." It was shown in prior work that adding emotional qualities to the robot increased interaction and desire from human participants to help the robot in learning tasks (McNeill and Kennington, 2020).

Planner When a participant utters a command to find a specific object, the planner enters a `searching` state where it takes random turns and drive distances while YOLOv11 checks each incoming camera frame for the target object. If the target is not detected after a specific time (20 seconds for our experiments), the robot displays frustration using the GRED Module. After 10 more seconds without a detection, the search reaches a `stuck` state. To establish the base capacity of the system, we compare two planners: a rule-based planner and a visual language model (VLM) based planner. Both planners use the YOLOv11 bounding box and object label information. The rule-based planner follows a set of rules to make random movements to explore the space until the specified object is found, and moves closer to the object once it is. The VLM-based planner uses a local LM to make a plan for robot actions (i.e., non-random) to move closer to a specified object.

The final experiment extends the Planner with a `stuck` state and `hint` state. After the `stuck` state is reached, the Planner enters the `hint` state and listens for speech from the participant. If a hint from the participant is given (e.g., *turn right* or *look behind you*), Cozmo executes the movement and checks YOLO for the target. If three consecutive hints do not result in the robot finding the object, the system enters the `stuck` state again. If the target is detected at any point during search or `hint` states, the Planner enters an `idle` state and Cozmo says *found it* while GRED generates and plays a `joy_hope` display on Cozmo. How the `stuck` signal is delivered differs by either a speech event where the robot utters *can't find*, or a *gaze trigger*, explained below.

Gaze Trigger In the gaze condition, the `stuck` state triggers a physical act rather than speech. Because Cozmo's screen eyes have limited expressivity, the cue is not eye movement alone: Cozmo rotates its whole body to face the participant, adjusts its head toward their face, and holds that facing orientation. This is made possible by the `Face Detection Module`, which rotates in place at startup until it detects the participant's face and stores that location in Cozmo's SLAM Module. When the `stuck` state triggers, Cozmo uses that stored position to turn toward the person and sets its head angle by distance so it is oriented at their face. Cozmo then holds that orientation for three seconds so the participant can register the facing cue. This pause is deliberate: prior work shows that establishing direct eye contact before any subsequent gaze shift changes how that cue is processed (Kompatsiari et al., 2018). After the hold, Cozmo remains facing the participant and does not turn back immediately. The head resets only when the next action begins, whether that is a hint movement or a new

search command from the participant, so the signal stays visible until the participant responds. The baseline condition produces no physical movement and no orientation change at the stuck moment, so the between-condition contrast is this embodied facing/gaze cue versus the verbal give-up.

4 Experiment 1: Planner Comparison

We first establish the base capability of the system given our task which requires tabletop robot navigation: a human participant tells the Cozmo robot to find a specific object type, then Cozmo autonomously searching for the targeted object on a tabletop under two visual planners that share the same robot platform, camera input, safety guards and completion criteria, and differ only in how each frame is interpreted to select the next action. Figure 4 shows the full system.

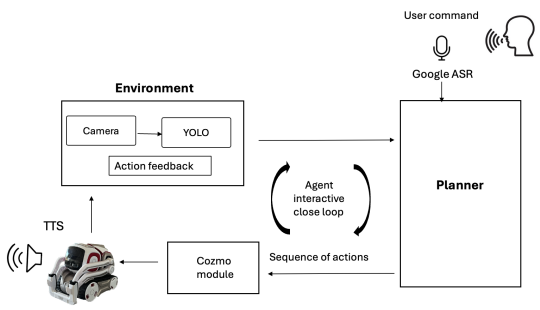


Figure 4: Base system architecture.

Planners. We first evaluate two different Planner variants. The two Planners differ only in how each frame is interpreted to select the next action. The **VLM-planner (Qwen2.5-VL-7B)** interprets the scene to perform object localization, spatial reasoning, and visibility assessment, and selects actions based on semantic scene understanding rather than geometry alone. The **rule-based planner** detects the target object in each frame and maps bounding-box geometry to navigation decisions through rule-based logic.

Procedure: We conducted a within-subjects study with ten participants, with Planner order randomized across participants. Each participant completed two 15-minute sessions, one for each Planner, in the same tabletop workspace with a fixed object set (car, bicycle, chair, and bottle). During each session, participants issued voice commands (e.g., *find the chair*) specifying the target object and observed Cozmo's behavior. Participants were allowed to reposition Cozmo and the objects between attempts.

After each session (a session could have several instances of the participant instructing Cozmo to find a specific object), participants rated the Planner along two dimensions: task performance (six custom 5-point Likert items assessing whether the robot's behavior matched the requested task) and perceived intelligence from the the Godspeed subscale (Bartneck et al., 2009), consisting of five 5-point semantic differential items.

Metrics. Planner performance is evaluated using objective and subjective measures. Objective measures, drawn from system logs, capture task success rate (whether the system found the object). Subjective measures from post-session questionnaires capture perceived performance through task performance and perceived intelligence ratings. Because participants set their own object arrangements on the table, some failures may result from scene difficulty rather than planner limitations. Subjective ratings therefore complement objective metrics by capturing aspects of behavior not reflected in completion outcomes. Statistical analyses follow the paired within-subjects design and small sample size. Participant-level comparisons use the Wilcoxon signed-rank test.

Results. The VLM planner found the target object in 41 of 47 attempts (87.2%), compared to 27 of 41 for rule-based (65.9%). This difference was significant (Wilcoxon signed-rank test, $p = 0.039$, $r = 0.79$), indicate that the VLM planner is more reliable at finding the target object than rule-based planner. Table 1 shows the detail results.

Table 1: Object-finding performance by system.

System	Found	Not found	Found rate
VLM planner	41 / 47	6 /47	87.2%
Rule-based planner	27 / 41	14 /41	65.9%

Subjective ratings followed the same pattern as objective performance. Participants rated the VLM planner higher on Task Performance (4.20 vs. 3.37; Wilcoxon signed-rank test, $p = 0.016$, $r = 0.76$) and Perceived Intelligence (4.02 vs. 3.30; $p = 0.016$, $r = 0.76$), with large effect sizes for both measures. (Appendix D, Figure 9 summarizes the ratings).

Discussion. The VLM-planner outperforms the rule-based planner on both objective and subjec-

tive measures, completing 87.2% of tasks versus 65.9% for the rule-based planner, with significantly higher ratings on task performance and perceived intelligence. In most of these cases in which the VLM planner found the target while the rule-based planner did not, participants created complex setups where the target was occluded or in an unusual pose (Figure 5), degrading YOLOv11's detection confidence. When detection confidence drops, the rule-based planner lacks sufficient information to act, since its decisions depend entirely on bounding-box features. The VLM planner remains effective in these cases by identifying the target from scene-level context.

Specifically, 34% of rule-based planner attempts ended without finding the target, so failure was common even though finding the target was still more frequent. This failure moment is what our experiment is built on: when the robot gets stuck, how should it signal that to the human?



Figure 5: Occluded object and unusual object pose case (images are extracted from Cozmo's camera view)

5 Experiment 2: Gaze vs. Verbal Stuck Signal

With our planner baseline established, we now turn to the main experiment. Our goal is to evaluate a system that uses the rule-based planner with the added capability of requesting help from the user through eye gaze and compare it to the VLM planner. Our hypothesis is that the rule-based planner, while not as effective as the VLM-planner on its own, will be as or more effective at the object finding task when it elicits help from humans using eye gaze.

Procedure. Our experiment uses a within-subjects design with two conditions: Baseline (verbally uttering *can't find*) and eye gaze (physical orientation toward the participant). Each participant completes both conditions in randomized order to control for order effects. Both sessions use the rule-based planner. The target objects are the same common household items as in the preliminary experiment (cup, chair, potted plant, bicycle) for which bounding-box detection is reliable. Figure 6 shows the tabletop workspace and a participant observing Cozmo during a session. Before each session, the experimenter places the target object out of Cozmo's initial camera field of view so the `stuck` state triggers reliably.

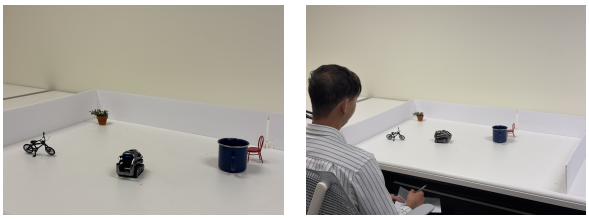


Figure 6: Left: tabletop workspace with Cozmo and target objects (cup, chair, potted plant, bicycle). Right: participant observing Cozmo, holding the paper sheet used to record found objects.

Participants are introduced to Cozmo with the following instructions: "In this session, you will be working alongside a Cozmo robot. Cozmo's goal is to explore the environment and find as many target objects as possible. Cozmo understands spoken English. If at some point it seems like it needs your help, you can try talking to it." No hint vocabulary is disclosed and no mention is made of the gaze condition (full instruction sheet in Appendix A). Participants are given a paper sheet to fill in during the session, listing which objects the robot found on its own and which it found with their help (Appendix B). As before, participants can set the objects where they please on the table. Each session covers four objects in fixed order and lasts approximately 20 minutes, followed by a post-session questionnaire containing custom Likert items assessing task performance and collaboration perception, and the Godspeed robot evaluation questionnaire (Bartneck et al., 2009). Total participant time is around one hour. Participants received a $12 Amazon gift card as compensation. Eight participants took part in the study.

Metrics & Baseline. Only the `stuck` state differs between conditions. *Baseline* says "can't find"; *Gaze* turns Cozmo's body to face the participant

and holds that orientation (a full-body facing cue, not eye motion alone). Mid-search timing and GRED `anger_frustration` displays are identical in both conditions, as are hint commands after the stuck cue; emotion is therefore not confounded with condition. The manipulated factor is only how failure is disclosed at the stuck moment. The primary measure is objects found with participant help, recorded on a paper notes sheet (Appendix B). Objects found alone and total objects are secondary measures.

After each session, participants completed a questionnaire (Appendix C) with eight custom 5-point Likert items on task experience and collaboration, and the Godspeed questionnaire (Bartneck et al., 2009) covering Anthropomorphism, Animacy, Perceived Intelligence, and Likability. The two custom items target our research question (i.e., *I felt invited to help* and *robot movements made sense when stuck*). All comparisons use the Wilcoxon signed-rank test with effect size r (exact two-tailed p, effect size $r = |z|/\sqrt{N_{\text{nontied}}}$).

Results. Behavioral outcomes: Participants recorded which objects Cozmo found on its own and which it found with their help on a paper sheet during each session.

	Baseline				Gaze			
P	Alone	Help	Total	Miss	Alone	Help	Total	Miss
1	3	1	4	0	3	1	4	0
2	3	1	4	0	2	2	4	0
3	3	0	3	1	3	1	4	0
4	3	0	3	1	2	0	2	2
5	2	0	2	2	3	1	4	0
6	4	0	4	0	3	1	4	0
7	1	0	1	3	2	1	3	1
8	1	0	1	3	2	2	4	0
Sum	20	2	22	10	20	9	29	3
Mean	2.50	0.25	2.75	1.25	2.50	1.13	3.63	0.38
SD	1.07	0.46	1.28	1.28	0.53	0.64	0.74	0.74

Table 2: Object counts per participant. Miss = objects never found in either condition. Help: $W = 0$, $p = 0.031$, $r = 0.899$.

Per-participant counts for found objects are in Table 2. Only 2 of 8 participants helped in Baseline; 7 of 8 helped in Gaze. Objects found with help: Gaze mean = 1.13 (SD = 0.64) vs. Baseline mean = 0.25 (SD = 0.46), $W = 0$, $p = 0.031$, $r = 0.899$. Objects found alone were the same in both conditions (mean = 2.50; $W = 10.5$, $p = 1.0$), showing the effect is specific to the collaboration channel. Total objects trended toward Gaze (3.63 vs. 2.75; $W = 1.5$, $p = 0.188$, $r = 0.724$) but did not reach significance at $N = 8$.

Qualitatively, the conditions looked different at the stuck moment. In Baseline, most participants watched after hearing "can't find" and did not speak; only two participants offered help across the whole condition, and helping was rare (2 objects found with help in total). In Gaze, after Cozmo turned to face them, seven of eight participants responded with spoken guidance, typically short directional hints such as *turn left*, *turn right*, or *look behind you*, which the planner could execute in the hint state. Objects found alone stayed the same across conditions (20 vs. 20), so the extra successes came from this help channel rather than from better solo search. We did not run a formal coding of every utterance; this account summarizes the pattern already visible in the per-participant counts in Table 2.

Gaze vs. VLM Planner: In the baseline setting, the robot was able to find 20 objects without help and an additional 2 with help (i.e., participants offered hints when the robot said *I'm stuck*). In the Gaze system where the robot looked at the participant when stuck, the robot was able to find 20 objects without help, and an additional 9 objects with help. The baseline system had 68.75% (22 out of 32) successfully found objects, whereas in the Gaze system, 87.87% (29 out of 33) resulted in found objects. Compared to the VLM Planner which had a 87.2% success rate, using the more basic planner with increased interaction using Gaze to elicit help enabled the robot to successfully find objects at a rate on-par (in fact, slightly better) than the vision-language model based planner. This supports our hypothesis that gaze with a simpler planner is comparable to a more sophisticated planner. This is a very positive result, as it shows that simple planning with elicited help from humans, which is akin to the behavior of children who seemingly randomly explore their environment yet look at caregivers for help when frustrated of confused, is an effective and collaborative method of exploration.

Custom Likert items. After each session, participants rated eight custom Likert items (1–5) on task experience and collaboration. Two questions are specific to our task: (Question 6: *I felt invited to participate when the robot seemed unsure what to do*) and (Question 9: *The robot's movements made sense for when it was unable to find an object*).

	Baseline		Gaze			
Item	M	SD	M	SD	W	p
Q3 Interesting	3.88	0.99	3.75	0.71	2.0	1.000
Q4 More time	4.00	1.20	4.13	1.36	2.0	1.000
Q5 Found object†	3.13	0.83	3.71	0.76	2.0	0.250
Q6 Felt invited*	**3.63**	**1.19**	**3.38**	**1.06**	**3.5**	**0.750**
Q7 Engaged	3.63	0.92	3.75	0.71	6.0	1.000
Q8 Movements/search†	4.00	0.76	3.71	0.76	2.5	0.625
Q9 Movements/stuck*	**3.88**	**0.99**	**3.63**	**1.06**	**1.5**	**0.750**
Q10 Collaboration	3.75	0.71	3.50	1.07	1.5	0.750

Table 3: Custom Likert means (1–5) and exact Wilcoxon statistics. **Bold** rows are primary items (*). † $N = 7$ (one missing response). No item reached significance ($\alpha = 0.05$).

No item reached significance (Table 3). Q6: Baseline = 3.63 (SD = 1.19), Gaze = 3.38 (SD = 1.06), $W = 3.5$, $p = 0.750$, $r = 0.274$. Q9: Baseline = 3.88 (SD = 0.99), Gaze = 3.63 (SD = 1.06), $W = 1.5$, $p = 0.750$, $r = 0.463$. Perceived collaboration (Q10) showed no difference either (Baseline = 3.75, Gaze = 3.50; $W = 1.5$, $p = 0.750$).

Two factors may explain these Likert results. First, gaze appears to have worked as an *implicit* cue rather than an explicit one: participants helped more in the gaze condition, which resulted in more objects found, yet did not rate themselves as feeling more invited or as better understanding the stuck signal. We conjecture that a facing/gaze cue can shift collaborative behavior without the observer consciously registering it as a request (Kompatsiari et al., 2018). Second, individual variation was high on both items (SD 1.06–1.19 on a 5-point scale), which reduces power at $N = 8$. Effect sizes ($r = 0.27$ for Q6, $r = 0.46$ for Q9) suggest a larger study would be needed to confirm these trends. See Appendix for Godspeed questionnaire results, which similarly showed no difference in human perceptions between the two systems.

The low helping rate in the verbal baseline also needs explanation. The phrase “can’t find” names failure, but it does not clearly invite participation; participants may have heard it as the end of a trial rather than a request for help. Robot speech can also be hard to hear or parse, so some people may simply have missed or underweighted the cue. We therefore cannot fully separate “gaze/orientation is strong” from “this particular verbal phrase is weak.” A stronger verbal topline would use an explicit invitation (e.g., “can you help me?” or “please tell me where to look”), as in prior work on robot help-seeking (Cha and Matarić, 2016). We did not run that condition here.

6 Limitations

This study has several limitations. First, both experiments were conducted with relatively small sample sizes (preliminary experiment: $n = 10$, main experiment: $n = 8$), which may limit statistical power, so results may reflect random variation rather than true differences; future studies should replicate these findings with larger, more diverse participant pools. Second, the GRED emotional expressions were not always clear for the participants to interpret, which likely weakened their usefulness as aid cues; we also did not run an emotion-off condition, so we cannot claim that emotion plus gaze is jointly necessary. This limitation could be mitigated by refining the GRED output for more consistent interpretation and by testing an emotion-off ablation in future work. Third, the setup in our experiment was simpler and did not include many failure cases where the baseline system struggled with object detection. Future studies should test more complex and varied environments. Additionally, our experiment evaluated only the rule-based planner, which has limited ability to interpret human-provided hints (e.g: "the cup is behind you"); future work could improve its language grounding or extend the evaluation to VLM-based planners. Finally, both experiments were conducted on the Cozmo platform, which limits the robot’s action space and expressive range; future work should validate these findings on more capable robotic systems.

7 Conclusions & Future Work

This paper examined whether gaze, a simple, non-verbal social signal, can help an embodied robot communicate failure to a human collaborator during a object-search task. Specifically, we asked whether this signal could help the robot succeed even when it lacks the reasoning needed to solve the problem on its own. We first compared two planners that differ in capability: a rule-based planner that reacts to raw detections, and a vision-language model planner that reasons over full scene semantics. The VLM planner found the target object far more often (87.2%) than the rule-based planner (65.9%), confirming that semantic reasoning helps when a search grows difficult.

This capability gap set up the contrast at the

center of our design. Rather than treating the rule-based planner as simply a weaker version of the VLM planner, we treat it as a stand-in for an agent that has not yet developed full semantic and linguistic reasoning, much like a young child who cannot yet utter request for help due to a lack of vocabulary, but who still succeeds by drawing a caregiver into helping through emotional displayes and eye gaze. This raised our central question: if the robot lacks that reasoning capability, can interaction with a human partner make up the difference instead? To test this, we added a gaze-based *stuck* signal to the rule-based planner and compared it against a verbal 'give-up' as the baseline, while holding mid-search GRED frustration displays constant across both conditions. We found that the rule-based planner reached 87.87% with gaze, matching the VLM planner outright, while the verbal baseline raised it only to 68.75%.

Gaze still worked even though participants did not always say they felt invited to help, which suggests it acted as a subtle cue rather than a direct request (Kompatsiari et al., 2018). Taken together, our results show that a robot does not need full linguistic or semantic capability to collaborate successfully with a human: interaction itself, prompted by a well-timed nonverbal signal, can recruit the help needed to make up for capability the robot does not have, in the same way that young children rely on gaze rather than language to get help from a caregiver (Plane et al., 2018).

For future work, we plan to test whether gaze will work with a VLM-planner that is trained to be more on par with child-level exploration in more visually cluttered environments, with a larger participant pool, and to compare gaze against an explicit verbal invitation baseline (e.g., "can you help me?") rather than only a failure announcement such as "can't find." Importantly, the interaction should result in the robot not just finding objects, but exploring them such that it can also learn grounded representations of the objects connected to referring expressions made by humans to them, all learned through interaction.

Acknowledgements. We are grateful to the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 2429968 and 2140642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Rista Baral, Bethany Grenz, and Casey Kennington. 2025. Recognizing and generating novel emotional behaviors on two robotic platforms. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21503–21510. IEEE.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphization of robotic agents. *International journal of social robotics*, 1(1):71–81.

Elizabeth Cha and Maja Matarić. 2016. Using nonverbal signals to request help during human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5070–5076. IEEE.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. 2018. On the role of eye contact in gaze cueing. *Scientific reports*, 8(1):17842.

Matteo Lavit Nicora, Pooja Prajod, Marta Mondellini, Giovanni Tauro, Rocco Vertechy, Elisabeth André, and Matteo Malosio. 2024. Gaze detection as a social cue to initiate natural human-robot collaboration in an assembly task. *Frontiers in Robotics and AI*, 11:1394379.

Anna Manaseryan, Porter Rigby, Brooke Matthews, Catherine Henry, Josue Torres-Fonseca, Ryan Whetten, Enoch Levandovsky, and Casey Kennington. 2025. rrSDS 2.0: Incremental, modular, distributed, multimodal spoken dialogue with robotic platforms. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Avignon, France. Association for Computational Linguistics.

David McNeill and Casey Kennington. 2019. Predicting human interpretations of affect and valence in a social robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany.

David McNeill and Casey Kennington. 2020. Learning word groundings from humans facilitated by robot emotional displays. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual. Association for Computational Linguistics.

Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):1–33.

Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68.

Jekaterina Novikova, Leon Watts, and Tetsunari Inamura. 2015. Emotionally expressive robot behavior improves human-robot collaboration. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 7–12. IEEE.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. 2018. Predicting perceived age: Both language ability and appearance are important. In *Proceedings of SigDial*.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

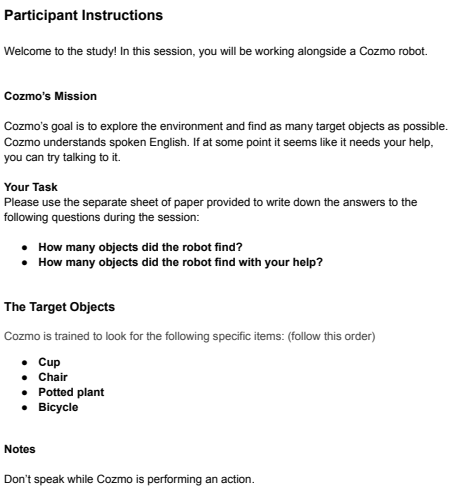
Ramtin Tabatabaei, Vassilis Kostakos, and Wafa Johal. 2025a. Gazing at failure: Investigating human gaze in response to robot failure in collaborative tasks. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 939–948. IEEE.

Ramtin Tabatabaei, Vassilis Kostakos, and Wafa Johal. 2025b. Real-time detection of robot failures using gaze dynamics in collaborative tasks. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1660–1664. IEEE.

Alexandra Weidemann and Nele Rußwinkel. 2021. The role of frustration in human–robot interaction–what is needed for a successful collaboration? *Frontiers in psychology*, 12:640186.

Shujie Zhou and Leimin Tian. 2020. Would you help a sad robot? influence of robots' emotional expressions on human-multi-robot collaboration. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*, pages 1243–1250. IEEE.

A Participant Instructions



Participant Instructions

Welcome to the study! In this session, you will be working alongside a Cozmo robot.

Cozmo's Mission

Cozmo's goal is to explore the environment and find as many target objects as possible. Cozmo understands spoken English. If at some point it seems like it needs your help, you can try talking to it.

Your Task
Please use the separate sheet of paper provided to write down the answers to the following questions during the session:

- **How many objects did the robot find?**
- **How many objects did the robot find with your help?**

The Target Objects

Cozmo is trained to look for the following specific items: (follow this order)

- **Cup**
- **Chair**
- **Potted plant**
- **Bicycle**

Notes

Don't speak while Cozmo is performing an action.

Figure 7: Full instruction sheet given to participants at the start of each session.

B Participant Notes Sheet

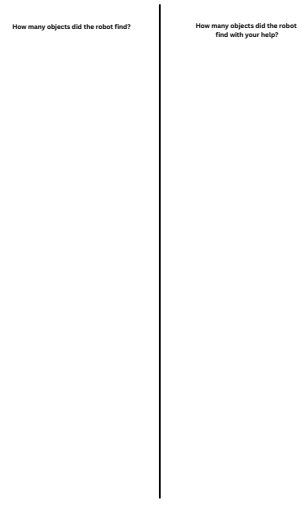


Figure 8: Paper sheet given to participants to record behavioral observations during the session.

C Post-Block Questionnaire

Participants completed the following questionnaire after each block. All rating items use a 5-point scale. Endpoint labels are shown in parentheses.

Part 1: Overall experience

1. How many objects did the robot find on its own? (numeric)
2. How many objects did the robot find with your help? (numeric)
3. How interesting was the robot to interact with? (1 = Not at all; 5 = Very)
4. Would you like to spend more time with the robot? (1 = Not at all; 5 = Very much)
5. Cozmo found the object I asked it to find. (1 = Strongly disagree; 5 = Strongly agree)
6. I felt invited to participate when the robot seemed unsure what to do. (1 = Strongly disagree; 5 = Strongly agree)
7. I felt engaged with the robot, not just watching it. (1 = Strongly disagree; 5 = Strongly agree)
8. The robot's movements made sense for finding the object. (1 = Strongly disagree; 5 = Strongly agree)
9. The robot's movements made sense for when the robot was unable to find an object. (1 = Strongly disagree; 5 = Strongly agree)
10. I felt we were working together to achieve a shared goal. (1 = Strongly disagree; 5 = Strongly agree)

Part 2: Impression of the robot — Godspeed questionnaire (Bartneck et al., 2009) *Anthropomorphism*

1. How aware did the robot seem? (1 = Unconscious; 5 = Conscious)
2. How natural did the robot seem? (1 = Artificial; 5 = Lifelike)
3. How would you describe the robot's movement? (1 = Moving rigidly; 5 = Moving elegantly)

Animacy

1. How alive did the robot appear? (1 = Dead; 5 = Alive)
2. How lively did the robot appear? (1 = Stagnant; 5 = Lively)
3. How interactive was the robot? (1 = Inert; 5 = Interactive)

Perceived Intelligence

1. How competent did the robot seem? (1 = Incompetent; 5 = Competent)
2. How knowledgeable did the robot seem? (1 = Ignorant; 5 = Knowledgeable)
3. How intelligent did the robot seem? (1 = Unintelligent; 5 = Intelligent)
4. How sensible were the robot's actions? (1 = Foolish; 5 = Sensible)

Likability

1. Please rate your impression of the robot in this session. (1 = Unlikable; 5 = Likable)
2. Please rate your impression of the robot in this session. (1 = Unpleasant; 5 = Pleasant)

C.1 Part 3: Godspeed results

Godspeed subscales We also administered the Godspeed questionnaire (Bartneck et al., 2009), which measures four dimensions of robot perception on 5-point bipolar scales. *Anthropomorphism* (3 items: Unconscious–Conscious, Artificial–Lifelike, Moving rigidly–Moving elegantly) captures how human-like the robot feels. *Animacy* (3 items: Dead–Alive, Stagnant–Lively, Inert–Interactive) captures how alive and active it appears. *Perceived Intelligence* (4 items: Incompetent–Competent, Ignorant–Knowledgeable, Unintelligent–Intelligent, Foolish–Sensible) captures how capable it seems. *Likability* (2 items: Unlikable–Likable, Unpleasant–Pleasant) captures how pleasant it is to interact with. Each subscale score is the mean of its items; if an item is missing for a participant, it is excluded from that participant's mean.

No subscale differed significantly (Table 4). Anthropomorphism: Baseline = 2.42 (SD = 1.17), Gaze = 2.14 (SD = 0.74), $W = 10.5$, $p = 1.0$ ($N = 7$). Animacy: Baseline = 3.25 (SD = 0.99), Gaze = 2.96 (SD = 0.84), $W = 6.0$, $p = 0.813$,

$r = 0.181$. Perceived Intelligence: Baseline = 3.16 (SD = 0.63), Gaze = 3.19 (SD = 0.70), $W = 14.0$, $p = 1.0$. Likability: Baseline = 4.31 (SD = 0.75), Gaze = 4.63 (SD = 0.52), $W = 1.5$, $p = 0.375$, $r = 0.639$.

This is expected. Godspeed reflects participants' overall impression of the robot across the full session, not just the stuck moment, so large between-condition differences are unlikely. Perceived Intelligence was nearly identical (3.16 vs. 3.19), showing that the gaze turn did not make Cozmo seem less capable. Anthropomorphism was low in both conditions, consistent with Cozmo being a small toy robot. Likability trended higher for Gaze ($r = 0.639$), the largest effect across all subscales, but did not reach significance at $N = 8$.

	Baseline		Gaze			
Subscale	M	SD	M	SD	W	p
Anthropomorphism†	2.42	1.17	2.14	0.74	10.5	1.000
Animacy	3.25	0.99	2.96	0.84	6.0	0.813
Perceived Intelligence	3.16	0.63	3.19	0.70	14.0	1.000
Likability	4.31	0.75	4.63	0.52	1.5	0.375

Table 4: Godspeed subscale means (1–5) and exact Wilcoxon statistics ($N = 8$ unless noted). † $N = 7$ for Gaze (P4 missing). Likability showed the largest effect ($r = 0.639$), favouring Gaze. No subscale reached significance ($\alpha = 0.05$).

D Figures

Figure 9 summarizes the ratings of participant on Task performance and Perceived Intelligence of the Base system between the VLM planner and Yolo planner (baseline).

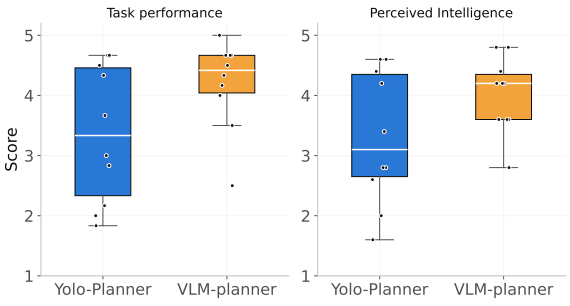


Figure 9: Task performance and Perceived Intelligence subscale by planner (Base system)

Speaker gaze and disfluencies during storytelling toward a robot or a human: The role of partner type and temperamental shyness

Nils F. Tolksdorf
Institute for German as a Foreign Language,
SMARTcognition Lab,
Heidelberg University
tolksdorf@idf.uni-heidelberg.de

Adriana Hanulíková
Institute for German as a Foreign Language,
SMARTcognition Lab,
Heidelberg University
hanulikova@idf.uni-heidelberg.de

Abstract

Storytelling in face-to-face dialogue involves coordinating multiple modalities, including speech and gaze. When the addressee is a social robot rather than a human, narrators' speech and gaze behavior may differ. However, few studies have examined both modalities within the same narratives, and it remains unclear how different aspects of multimodal behavior vary depending on whether the interaction partner is a social robot or a human. In this study, we examined how adults narrated picture-based stories to a social robot and a human interlocutor. Two modalities of narrative production were considered: pausing, separated into filled and unfilled pauses, and gaze behavior during the narration, directed either toward the story pictures or toward the listener. Critically, we examined whether shyness, considered as a continuous dimension of individual variation, relates to pausing and gaze behavior. Results revealed that speakers produced fewer filled pauses when narrating to the robot than to the human, while unfilled pauses did not differ. Additionally, higher shyness was associated with longer, though not more frequent, filled pauses in human-directed narration, although shyness did not reliably predict the magnitude of the partner effect. We discuss how our findings may contribute to a more nuanced understanding of multimodal interactions with embodied agents.

1 Introduction

Human communicative interaction in face-to-face settings is characterized as a multimodal process (Levinson, 2016). When telling a story, for example, speakers produce speech that includes disfluencies, such as filled and unfilled pauses, while also coordinating their gaze between the story pictures and their interlocutor. Importantly, these modalities are adjusted to the addressee (Holler and Levinson, 2019; Hanulíková et al., 2026). Speakers have been shown to tailor their behavior to a model of their partner, and communicative behaviors directed to artificial partners such as social robots appear to form a distinct register with features shared with computer- and child-directed speech (Fischer et al., 2011; Jacka et al., 2025; Hanulíková et al., 2026). The characteristics of speech and gaze that a robot elicits may therefore reveal how speakers perceive their partner, which can in turn inform the design of systems that interpret and produce human multimodal behavior more contingently (Hessels and Fang, 2026). Although research comparing interactions with robotic and human partners has grown (e.g., Demir-Lira et al., 2020; Vollmer et al., 2018), less is known about speakers' multimodal behavior in communicative tasks, such as narrative settings towards a social robot compared to human-human interaction, as well as the direction of potential adjustments a person might take when interacting with a robotic partner. On the one hand, a robot may be perceived as less socially evaluative than a human partner (Tolksdorf et al., 2026) and as a machine that requires precise, unambiguous input. In this case, speakers may hesitate less, and looking at the robot's face may carry less social cost. On the other hand, interpersonal interaction is a co-constructed achievement in which partners monitor their addressees and rely on their feedback (Clark and Krych, 2004). A robotic interaction partner that is less familiar and provides fewer listener cues could lead to more disfluent production (Tolksdorf et al., 2021a) and potentially decrease the reciprocity of partner-directed gaze. Crucially, temperamental traits can decisively influence how a person coordinates social interactions. Yet, the influence of such individual differences (e.g., shyness) on multimodal behavior remains understudied. As a prevalent phenomenon and continuous temperamental dimension involving heightened attention to the social evaluative aspects of interaction (Sosa-Hernandez et al., 2023; Viertel and Tolksdorf, 2024), shyness may relate to how

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 117–126
October 29, 2026 ©2026 Association for Computational Linguistics

speakers regulate pausing and gaze directed at their partner. While previous research on the influence of shyness on multimodal behavior in human-robot interaction has primarily focused on children (e.g., Tolksdorf et al., 2021b; Tykhonenko et al., 2026), little research has examined this topic in adults. The present pilot study addresses these gaps by comparing filled and unfilled pausing and gaze allocation within the same narrations, when interacting with either a social robot or a human partner, and by relating both modalities to individual levels of shyness. The aims of this work are as follows:

- To investigate whether filled and unfilled pauses, as well as gaze behavior directed toward story pictures and interaction partners, vary when a picture-based story is narrated to a robot versus a human interaction partner.
- To examine whether individual differences in shyness are associated with variation in hesitation and gaze behavior.

2 Related Work

2.1 Speaking to a robot versus a human partner

A body of research has shown that speakers may adapt aspects of their speech to their addressees, depending on partner representations, communicative demands, and the interactional context (Fischer, 2016; Brennan and Hanna, 2009; Ferreira, 2019). However, although previous studies have documented linguistic and prosodic adaptations to different human addressees, comparatively little is known about differences in disfluency and gaze related to the partner during semi-naturalistic narrative production with embodied artificial partners, such as social robots. In particular, speech to artificial partners is reported to share features with other partner-adapted registers (Fischer et al., 2011). In addition, prior research suggests that backchannels and mutual monitoring in human–human interaction can support fluent production (Clark and Krych, 2004). Because robotic systems typically provide fewer such cues, narration to a robot may be comparatively more disfluent, consistent with earlier findings that children involved their caregiver more, and that the interactional flow was more often disrupted with a robot than with a human partner (Serholt, 2018). At the same time, speakers may treat a machine partner as requiring precise, unambiguous input (Pang et al., 2025), and accordingly produce more controlled, less disfluent speech. Given that these findings may predict opposite directions, we leave the direction of partner effects open a priori.

2.2 Disfluencies as cognitive-communicative signals

Filled and unfilled pauses can serve multiple functions in interaction. Clark and Fox Tree (2002) argued that markers such as *uh* and *uhm* can act as conventionalized signals of an upcoming delay in turn-taking, and prior work converges on a dual role for these fillers in speech planning and listener-oriented signalling (Corley and Stewart, 2008). Unfilled pauses, by contrast, have been more closely linked to cognitive planning (Zellner, 1994). If speakers adapt their listener-directed behavior primarily to the type of interaction partner, then partner effects may be more likely for filled pauses than for unfilled ones. However, if reduced listener feedback leads to increased disruption in speech production, then both pause types may be affected. We therefore analyse filled and unfilled pauses separately and ask whether partner type and speaker characteristics act differently on each.

2.3 Gaze during storytelling

Speaker gaze in conversation has been described as serving regulatory, monitoring, and expressive functions (Kendon, 1967). In storytelling settings, narrators' gaze toward recipients has been associated with the elicitation of listener feedback (Bavelas et al., 2002), although more recent eye-tracking evidence suggests that this association is less uniform than initially proposed (Zima, 2020). Gaze toward interaction partners during hesitations can also invite collaboration, for instance when searching for a word (Auer and Zima, 2026), whereas gaze aversion has been observed during hesitation phases and before complex segments of talk (Doherty-Sneddon and Phelps, 2005; Auer and Zima, 2026). In line with this notion, gaze directed at a partner primarily serves to elicit feedback. A robot offering fewer cues may attract fewer initiations of gaze from its partner. Additionally, looking at a human's face may carry more social-evaluative weight, whereas gazing at a robot's face may be less costly and thus occur more frequently. Importantly, most existing evidence stems from face-to-face conversations without task-related visual materials. In picture-elicited narration, speakers can be expected to direct much of their gaze at

the pictures, so that partner-directed gaze becomes a comparatively infrequent behavior whose distribution across different partner types is not well documented (Hessels and Fang, 2026).

2.4 Shyness and socially evaluative influences on production

Temperamental shyness is characterized by high sensitivity to novel situational factors, including social aspects such as the concern about being evaluated negatively (Viertel and Tolksdorf, 2024). In this respect, social robots have been discussed as potential partners that can lessen the evaluative load within interaction (Peter and van Straten, 2024). The available evidence is, however, nuanced. In a dyadic peer paradigm investigating temporal coordination during interaction, conversational response times were related to children's observed social engagement, but not their self-reported shyness (Wilson et al., 2024). In adults with social anxiety disorder, a clinical construct distinct from temperamental shyness, the proportion of hesitation disfluencies was similar to that of controls, though silent pausing was more prevalent (Silber-Varod et al., 2016). Evidence on the influences of shyness on multimodal behavior in human-robot interaction largely stems from children and modalities such as proxemics, expressiveness, and response latencies (e.g., Tolksdorf et al., 2021b; Tykhonenko et al., 2026). Taken together, these findings can motivate the expectation that shyness, if it relates to production, may do so selectively — for pausing, possibly more in unfilled than in filled pauses. For gaze, partner-directed gaze is one of the more socially exposed behaviors in a narration task and may therefore be a plausible source of variation related to shyness. Because directly comparable evidence is limited, we examine this question without making a strong directional prediction. Since the social-evaluative demands of a robot and a human partner differ, shyness-related patterns may be partner-dependent.

2.5 The present study

This work is part of a broader project examining how speech production (e.g., lexical retrieval) in interaction with a robot compares with human–human interaction and depending on language background. Language background is a focus of the broader project, and is addressed here only exploratorily because of its potential relevance to hesitation behavior (Section 4.6). In the present study, adults narrated picture-based stories to a social robot and to a human partner. We focus on examining whether the characteristics of the interaction partner (human versus robot) are associated with differences in hesitation behavior and gaze behavior during narration when overt listener feedback is minimized. Additionally, we investigate whether temperamental shyness is associated with disfluencies and partner-directed gaze and whether such associations appear selective for the type of social partner (robot versus human).

3 Method

3.1 Participants

24 adults (19 female, 5 male; mean age 26.98 years, $SD = 6.2$) participated in the pilot study. 13 were first-language (L1) speakers of German and 11 were speakers with German as a second language (L2). Due to technical errors with the robot, two participants completed only the human-directed condition and are excluded from within-participant partner contrasts.

3.2 Partner conditions

In the robot condition, participants narrated to the Navel social robot (Toussaint et al., 2023). The robot's verbal behavior was generated via its built-in speech-to-text API and LLM integration, prompt-engineered to follow a standardized narrative-elicitation protocol (Gagarina et al., 2019): it invited the participant to tell the story, did not interrupt during the narration, and thanked the participant at the end. Both conditions were designed to be as comparable as possible, and during the narration, the robot remained in a stationary listening posture with its gaze directed ahead (see Figure 1A). The robot did not provide contingent verbal or nonverbal backchannel signals. In the human condition, participants narrated to a human experimenter whose listening behavior remained consistent as well. The experimenter did not initiate or return eye contact or produce backchannel signals and looked at the story pictures during the narration. Thus, both partners provided minimal, non-contingent listening behavior. This pattern was a deliberate design decision. Since the robot could not produce contingent listener responses, an unconstrained human listener might introduce differences in feedback, which would make it difficult to distinguish the effects of partner type from those of feedback. Therefore, holding listener responsive-

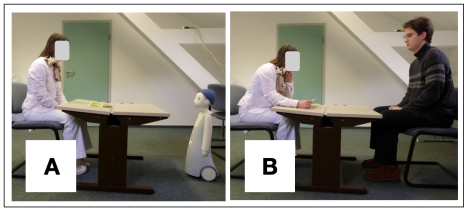


Figure 1: Experimental setup for robot- or human-directed storytelling interactions. (A) robot condition, (B) human condition.

ness constant at a minimal level may help to discern effects attributable to the partner itself rather than to the feedback it provides. However, this approach comes at the cost of naturalness. A listener who withholds backchannels and eye contact diverges from typical conversational storytelling, where listeners respond audibly and visibly (Bavelas et al., 2002; Zima, 2020).

3.3 Procedure and materials

Participants completed a single laboratory session. Narratives were elicited with the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2019), using parallel six-picture stories. Following the MAIN protocol, participants selected one of three envelopes, a procedure intended to create the impression that the partner did not already know the chosen story, and then narrated the story while viewing the picture sequence, which remained visible throughout. Narrations were video-recorded, and the duration of each narration was measured to calculate rates per minute. The order and assignment of stories were balanced across social partners, as was the order in which participants told stories to the robot or the human.

3.4 Disfluency and gaze coding

Disfluencies were annotated as filled pauses (e.g., *uh*, *uhm*) and unfilled (silent) pauses within the narration. Each pause was annotated with its onset and offset. From this annotation, we derived two measures per narration. First, as frequency measures, we computed filled and unfilled pauses per minute by dividing each participant's pause counts by the corresponding narration duration. Second, we calculated the mean duration of pauses for each participant and partner condition. Gaze was annotated manually, frame by frame, from the video recordings in ELAN (Wittenburg et al., 2006). Gaze targets are coded as (a) the picture sequence and (b) the partner's face region (the robot's face region in the robot condition). From this annotation we derive, per narration, the number of partner-directed gaze events per minute. We report event rates because they indicate how often speakers look at each target, which most directly corresponds to the gaze-initiation behavior that motivated our research questions. In the present picture-elicited narration, speakers direct most of their gaze toward the stimulus material. Consequently, partner-directed looks occur as brief departures from the pictures. Therefore, their frequency provides insight into how often the speaker turns toward the partner, whereas duration-based measures are dominated by the picture-directed baseline. Additionally, rates rather than raw counts were used because narration durations varied across participants. All coding was performed by a single trained coder.

3.5 Assessment of shyness

Temperamental shyness was assessed using the "Shyness and Sociability Scales for Adults" (Asendorpf, 1997; Asendorpf and Wilpers, 1998). This standardized self-report questionnaire was administered at the end of the session and provided a continuous score that was used for analyses.

4 Preliminary Results

4.1 Analytic approach

Pause rates (filled and unfilled pauses per minute) and gaze-event rates per minute were analyzed using nonparametric, rank-based ANOVA-type statistics (ATS; Brunner et al., 2018; Noguchi et al., 2012), with partner (robot, human) as a within-participant factor. For each comparison, we report the ANOVA-type statistic (ATS), its associated F-approximation (for which the denominator degrees of freedom are set to infinity), and the relative treatment effects (RTEs) for each condition. An RTE above .50 indicates that observations from that condition tend to take higher ranks in the pooled sample. The same rank-based approach was applied to participants' mean filled- and unfilled-pause durations in each condition. For the analysis of pause *durations*, duration values are defined only for participants who produced at least one pause of the relevant type in a given condition; participants without such pauses were excluded from the corresponding duration analysis (18 pairs for filled and 14 pairs for unfilled pauses). These analyses therefore concern mean pause duration among participants who produced at least one pause of the relevant type in

both conditions. Additionally, two further participants could not be included in the gaze coding, so gaze analyses are based on 20 complete pairs.

Associations with shyness were examined in two ways. First, we report Spearman rank correlations within each partner condition, describing how shyness relates to behavior within a partner condition. Additionally, to directly examine whether the magnitude of the partner difference varied with shyness, we calculated the within-participant difference between the human and robot conditions for each measure and correlated these difference scores with shyness. A significant association would indicate that the magnitude of the difference between partner conditions varied systematically with individual differences in shyness. Given the small sample size and the distributional characteristics of the behavioural measures, we used a rank-based analysis of within-participant difference scores as a parsimonious approach to examining whether the magnitude of partner differences varied with shyness. All shyness analyses are exploratory, and we did not apply correction for multiple testing, so individual p-values are not interpreted as confirmatory evidence. To examine whether partner effects differed by language background, we additionally included language background as a between-participant factor in the rank-based ATS model, with partner as a within-participant factor for filled and unfilled pauses.

4.2 Descriptive overview

Table 1: Descriptive overview per partner condition (medians with interquartile ranges). Medians are reported because the analyses are rank-based and the distributions are skewed.

	Human	Robot
Narration duration (min)	1.07 [0.92–1.57]	1.12 [0.85–1.37]
Pauses per minute		
Filled	8.26 [4.72–13.55]	3.15 [1.23–10.13]
Unfilled	1.97 [0.69–3.44]	1.74 [0.87–2.81]
Mean pause duration (s)		
Filled	0.79 [0.60–0.96]	0.70 [0.59–0.85]
Unfilled	1.13 [0.91–1.55]	1.05 [0.72–1.84]
Gaze events per minute		
Partner-directed	1.04 [0.29–2.18]	3.02 [0.29–6.38]
Picture-directed	1.13 [0.85–2.75]	3.44 [0.95–6.71]

4.3 Partner effects on pausing

Figure 2 shows pause rates in both conditions. Filled-pause rates were lower in robot-directed than in human-directed narration, and the ATS indicated a difference between conditions, $F(1, \infty) = 18.88$, $p < .001$, with relative treatment effects of $RTE = .58$ (human) and .42 (robot). Unfilled-pause rates did not differ between conditions, $F(1, \infty) = 0.04$, $p = .848$, $RTE = .51$ (human) and .49 (robot). Together, in this sample the partner effect appeared to be confined to filled pauses while the data provide no comparable indication for unfilled pauses. Pause durations showed no partner differences, for filled pauses, $F(1, \infty) = 0.93$, $p = .335$, $RTE = .53$ (human) vs. .47 (robot), or for unfilled pauses, $F(1, \infty) = 1.39$, $p = .239$, $RTE = .48$ (human) vs. .52 (robot).

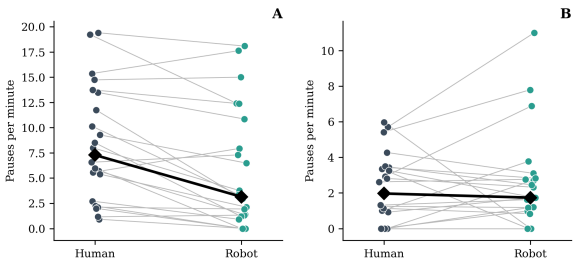


Figure 2: Filled (A) and unfilled (B) pause rates per participant and partner condition. Grey lines connect each participant's two narrations; black line marks the condition medians.

4.4 Partner effects on gaze

To investigate partner-directed gaze-event rates per participant and condition, we ran another ATS that did not indicate a significant difference, $F(1, \infty) = 2.46$, $p = .117$, $RTE = .45$ (human) and .55 (robot), although median rates were descriptively higher toward the robot ($Mdn = 3.02$ vs. 1.04 events per minute). Picture-directed gaze events were more frequent in robot-directed than in human-directed narration; the ATS indicated a difference between conditions, $F(1, \infty) = 3.94$, $p = .047$, $RTE = .44$ (human) and .56 (robot). Descriptively, gaze-event rates were higher in the robot condition for both targets. However, only the picture-directed comparison yielded a nominally significant difference.

4.5 Role of shyness

Table 2 reports Spearman correlations between shyness and the disfluency and gaze measures within each partner condition, together with the correlation between shyness and the within-participant difference between conditions.

Within the human condition, exploratory analyses suggested a possible association between higher

Table 2: Spearman correlations between shyness and each measure, per partner condition and for the within-participant difference (Δ = Human − Robot). p-values are uncorrected for multiple testing and are reported as exploratory.

	Human		Robot		Δ	
	r_s	p	r_s	p	r_s	p
Pauses per minute						
Filled	+.08	.696	−.10	.670	+.26	.235
Unfilled	+.18	.399	+.11	.626	−.25	.269
Mean pause duration (s)						
Filled	+.41	.046	+.13	.597	+.25	.318
Unfilled	−.27	.285	−.14	.602	+.11	.702
Gaze events per minute						
Partner-directed	+.03	.914	−.30	.203	+.50	.026

shyness and longer mean filled-pause durations, $r_s = +.41$, $p = .046$, whereas the corresponding robot-condition coefficient was weaker, $r_s = +.13$, $p = .597$. As illustrated in Figure 3, the difference-score test, however, did not indicate that shyness predicts the magnitude of the partner effect on filled-pause duration, $r_s = +.25$, $p = .318$. The within-condition pattern therefore describes an association in human-directed narration. Because the difference-score test was not significant, the present data provide no evidence that shyness relates to pausing differently depending on partner type.

For partner-directed gaze, the difference score was nominally associated with shyness, $r_s = +.50$, $p = .026$. Because the difference score was defined as human − robot, and partner-directed gaze was descriptively more frequent toward the robot, this positive association indicates that the relative increase in partner-directed gaze in the robot condition was smaller among participants with higher shyness scores. However, a comparable coefficient was obtained for picture-directed gaze ($r_s = +.48$, $p = .032$), which is consistent with a general difference in gaze shifting rather than a partner-specific effect. We report it as an exploratory observation for future testing.

These associations should further be considered in light of the linguistic heterogeneity of the sample. Shyness scores were comparable between L1 and L2 speakers ($M = 14.69$, $SD = 4.01$ vs. $M = 12.18$, $SD = 3.46$; $p = .110$), and as a relatively stable temperamental characteristic that can manifest across linguistic contexts (Viertel and Tolksdorf, 2024), shyness may not be expected to relate to pausing in a manner specific to L1 or L2 speech. Given the small sample, language background nevertheless remains a potential source of heterogeneity that should be examined more systematically in future work.

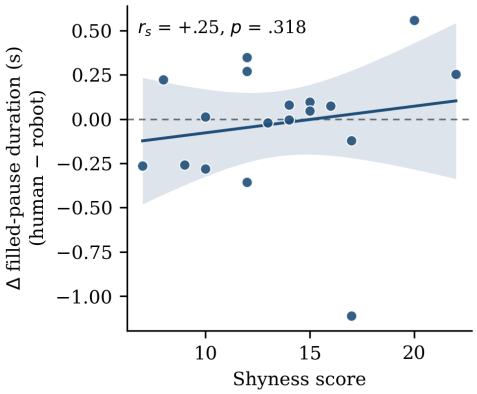


Figure 3: Association between shyness and the within-participant difference in mean filled-pause duration between conditions (human − robot). Positive values indicate longer filled pauses toward the human partner. The dashed line marks no partner difference; the shaded band is the 95% confidence interval of the linear fit.

4.6 Language background

For filled-pause rates, the ATS indicated a nominal L2 main effect, $F(1, 14.51) = 4.79$, $p = .046$, with L2 speakers at relatively higher filled-pause ranks overall, and an L2 × Partner interaction, $F(1, \infty) = 5.44$, $p = .020$: the reduction of filled pauses toward the robot was descriptively concentrated among L1 speakers (RTE human vs. robot: .50 vs. .29; median within-participant difference 3.27 pauses per minute), whereas the difference was smaller among L2 speakers ($RTE = .68$ vs. .61; median difference 1.37). For unfilled-pause rates, no effect involving language background approached significance. Interestingly, the reduction in filled pauses when speaking to the robot was evident in both language groups, and was descriptively more pronounced among L1 speakers. Exploratory analyses suggested that language background moderated the magnitude of partner differences in filled-pause rates, with larger human-robot differences observed among L1 speakers than L2 speakers. However, because subgroup sizes were small and no within-group tests were conducted, this pattern should be interpreted cautiously. Therefore, the main partner effect should not be assumed to be uniform across language backgrounds.

5 Discussion and Limitations

In this paper, we describe a pilot study that investigated the effects of different social partners (robot vs. human) and levels of shyness on multimodal narrative production with respect to disfluencies and gaze behavior. We found that robot-directed narrations included fewer filled pauses than human-directed narrations, whereas unfilled-pause rates, mean pause durations, and partner-directed gaze showed no significant differences. In the present interactional setting, both partners acted as minimally responsive listeners. Neither produced backchannels nor initiated or returned eye contact. Because overt listener feedback was minimised in both conditions, differences in overt contingent feedback are unlikely to be the primary explanation. However, the present design does not isolate partner expectations from other differences associated with interacting with a human versus a robot. For pausing, the observed pattern appears not to support the expectation that reduced listener support increases disfluency. Instead, fewer filled pauses toward the robot may converge with findings of lower disfluency rates in human–computer interaction and with evidence that speakers adapt to perceived artificial-partner capabilities (Branigan et al., 2011). Analysis of participants' gazes toward the robot or human revealed no differences, although partner-directed gazes were descriptively more frequent toward the robot. Because neither interaction partner returned the participant's gaze, mutual gaze was not established following partner-directed glances in either condition. We explain these findings with the possibility that glances toward the robot may partly reflect a visual inspection of a novel embodied partner rather than communicative signaling (Admoni and Scassellati, 2017; Hessels and Fang, 2026). This explanation is perhaps corroborated by our observations that gaze event rates also increased for the story pictures in the robot condition; thus, the pattern may reflect more frequent gaze shifting overall. However, the current event-based gaze measure cannot distinguish between more looking and more switching. A duration-based measure could differentiate between the two: the proportion of narration time spent on each target indicates the overall allocation of visual attention, while the mean time spent looking at each target indicates how long attention is sustained once directed there. A high event rate combined with short dwell times would indicate frequent redirection of gaze. In contrast, a high event rate with comparable dwell times would indicate more looking overall. We regard this as an informative next step for the gaze data. Moreover, higher shyness was associated with longer, not more frequent, filled pauses in human-directed narration. However, a direct test of whether shyness predicts the magnitude of the partner effect did not reach significance, so the present data do not establish that this association is specific to the human partner. The observed pattern raises the possibility that variation related to shyness may be more apparent in the temporal characteristics of pausing than in the pause frequency itself. Whether this variation appears to be partner-dependent, potentially because participants experience the partners differently in social or evaluative demands (Tolksdorf et al., 2026), remains an open question for future research.

While our study has taken initial steps to explore the effects of different social partners and shyness on multimodal behavior, several limitations need to be recognized. The sample is small and was gender-imbalanced. In addition, the human partner was a male experimenter, thus evaluative and gaze dynamics may vary with the gender composition of a dyad. A further limitation is that all disfluency and gaze annotation was performed by a single trained coder, so no inter-rater reliability estimate is available. Double-coding a subset of the recordings and reporting agreement is necessary to establish the reliability of these measures, and we regard this as a requirement for the larger confirmatory study. Another limitation concerns the extent to which the human condition can generalize. Because the human listener was limited to minimal, non-contingent behavior, our results characterize how speakers may adjust their behavior based on the type of partner when listener responsiveness is held constant. These results should not be interpreted as a full representation of natural human-to-human storytelling, in which the listener's responses are available and have been argued to encourage fluency (Clark and Krych, 2004). Although in the present design overt listener feedback was minimised to increase comparability across conditions, the absence of feedback may not have been perceived equivalently, and whether the observed partner effect would change with a responsive human listener cannot be determined from these data. Fluency-supporting feedback might reduce filled pauses toward the human listener, thereby attenuating the difference, whereas an unresponsive listener might have elevated hesita-

tions. A minimally responsive human may violate expectations for social interaction, whereas limited responsiveness may be more consistent with expectations for a social robot (Tolksdorf et al., 2026). In this case, the difference would be smaller under naturalistic conditions. Establishing this would require a responsive human baseline alongside the matched conditions reported here, which we consider a priority for future work. Moreover, each participant met the robot in a single session, so our results reflect first encounters; shyness- and novelty-related effects have been shown to change with familiarization (Tykhonenko et al., 2026).

6 Conclusions

In sum, this pilot study examined how adults' pausing and gaze behavior during picture-based storytelling varies with the type of social partner and with temperamental shyness. In this sample, narrations directed to the robot contained fewer filled pauses than those directed to the human partner, whereas unfilled pauses, pause durations, and partner-directed gaze showed no differences. Higher shyness was associated with longer, though not more frequent, filled pauses in human-directed narration, however, a test of the difference between conditions provided no evidence for a partner-dependent association. Additionally, in the present analysis, pause rates were normalized by narration duration, but not by more nuanced linguistic output. Two speakers may produce the same number of pauses per minute, yet differ substantially in speaking rate and amount of speech. Future analyses will therefore control for pauses by words or syllables and include speaking rate. Furthermore, future work will take into consideration gaze proportions and the temporal dynamics of gaze initiations, and how partner-directed glances relate to the timing of disfluencies.

References

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25–36.

Jens B. Asendorpf. 1997. *Schüchternheits- und Geselligkeitsskalen für Erwachsene*. Humboldt-Universität, Institut für Psychologie, Berlin.

Jens B. Asendorpf and S Wilpers. 1998. Personality effects on social relationships. *Journal of Per- sonality and Social Psychology*, 74:1531–1544.

Peter Auer and Elisabeth Zima. 2026. Does gaze aversion index dispreference or complexity of upcoming answers to questions? *Frontiers in Communication*, 11:1731835.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, 52(3):566–580.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57.

Susan E. Brennan and Joy E. Hanna. 2009. Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, 1(2):274–291.

Edgar Brunner, Arne C. Bathke, and Frank Konietschke. 2018. *Rank and pseudo-rank procedures for independent observations in factorial designs: using r and sas*. Springer Series in Statistics. Springer International Publishing, Cham.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Martin Corley and Oliver W. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*. *Language and Linguistics Compass*, 2(4):589–602.

Ö. Ece Demir-Lira, Junko Kanero, Cansu Oranç, Sümeyye Koşkulu, Idil Franko, Tilbe Göksun, and Aylin C. Küntay. 2020. L2 vocabulary teaching by social robots: the role of gestures and on-screen cues as scaffolds. *Frontiers in Education*, 5:599636.

Gwyneth Doherty-Sneddon and Fiona G. Phelps. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & cognition*, 33(4):727–733.

Victor S. Ferreira. 2019. A Mechanistic Framework for Explaining Audience Design in Language Production. *Annual Review of Psychology*, 70(1):29–51.

Kerstin Fischer. 2016. *Designing Speech for a Recipient: The roles of partner modeling, alignment and feedback in so-called 'simplified registers'*, volume 270 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company, Amsterdam.

Kerstin Fischer, Kilian Foth, Katharina J. Rohlfing, and Britta Wrede. 2011. Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12(1):134–161.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: multilingual assessment instrument for narratives – revised. *ZAS Papers in Linguistics*, 63:20.

Adriana Hanulíková, Nils Frederik Tolksdorf, and Sarah Kapp. 2026. Robot speech: how variability matters for child–robot interactions. *Frontiers in Robotics and AI*, 12:1725423.

Roy S. Hessels and Yu Fang. 2026. A visual perceptual perspective on gaze in social robotics. *Psychonomic Bulletin & Review*, 33(4):131.

Judith Holler and Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Rhys Jacka, Paola Raquel Peña, Sophie J Leonard, Éva Szekély, and Benjamin R. Cowan. 2025. Talking to...uh...um...machines: the impact of disfluent speech agents on partner models and perspective taking. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–7, Waterloo ON Canada. ACM.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Stephen C. Levinson. 2016. Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14.

K Noguchi, Y. R. Gel, E Brunner, and F Konietschke. 2012. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, 50(12):1–23.

Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. Does the appearance of autonomous conversational robots affect user spoken behaviors in real-world conference interactions? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, Yokohama Japan. ACM.

Jochen Peter and Caroline L. Van van Straten. 2024. Social robots and children: a field in development. In Leopoldina Fortunati and Autumn Edwards, editors, *The De Gruyter Handbook of Robots in Society and Culture*, pages 371–388. De Gruyter.

Sofia Serholt. 2018. Breakdowns in children's interactions with a robotic tutor: A longitudinal study. *Computers in Human Behavior*, 81:250–264.

Vered Silber-Varod, Hamutal Kreiner, Ronen Lovett, Yossi Levi-Belz, and Noam Amir. 2016. Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. In *Speech Prosody 2016*, pages 1211–1215. ISCA.

Linda Sosa-Hernandez, Charlotte Aitken, Savannah Heintzman, Elizabeth S. Nilsen, and Heather A. Henderson. 2023. Temperamental shyness and children's communicative behaviours in a goal-oriented task: A dyadic perspective. *Journal of Applied Developmental Psychology*, 87:101566.

Nils F. Tolksdorf, Camilla E. Crawshaw, and Katharina J. Rohlfing. 2021a. Comparing the effects of a different social partner (social robot vs. human) on children's social referencing in interaction. *Frontiers in Education*, 5(569615):1–12.

Nils F. Tolksdorf, Katharina J. Rohlfing, Angela Grimminger, and Kazuki Sekine. 2026. Social robots as a method for cross-cultural gesture elicitation in children. *Frontiers in Communication*, 11:1651149.

Nils F. Tolksdorf, Fanziska E. Viertel, Camilla E. Crawshaw, and Katharina J. Rohlfing. 2021b. Do shy children keep more distance from a social robot? Exploring shy children's proxemics with a social robot or a human. In *Interaction Design and Children*, pages 527–531, Athens Greece. ACM.

Claude Toussaint, Philipp T Schwarz, and Markus Petermann. 2023. Navel - a social robot with verbal and nonverbal communication skills. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, Hamburg Germany. ACM.

Valeriia Tykhonenko, Nils F. Tolksdorf, and Katharina J. Rohlfing. 2026. The relation between children's shyness and their contingent dialogical actions when reacting to a social robot's instructions. *Philosophical Transactions B*, 381(1943):20240375.

Fanziska E. Viertel and N. F. Tolksdorf. 2024. Multimodal communicative behaviours in shy children in assessment situations and social evaluative contexts. In Ozturk Nesrin, editor, *Contemporary Issues in Early Childhood Education and Care*, volume 27. IntechOpen, London.

Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. 2018. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, 3(21):eaat7111.

McLennon Wilson, Adrienne Richter Powell, Linda Sosa Hernandez, Emma Green, Claudia Labahn, and Heather Henderson. 2024. Shyness, social engagement, and conversational response times in children's dyadic interactions with an unfamiliar peer. *Social Development*, 33(3):e12734.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, page 5th.

B Zellner. 1994. Pauses and the temporal structure of speech. In E Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. Chichester.

Elisabeth Zima. 2020. Gaze and recipient feedback in triadic storytelling activities. *Discourse Processes*, 57(9):725–748.

FriendBench: Benchmarking Dyadic Familiarity Inference in Humans and Multimodal Large Language Models

Jeffrey M. Girard, Jason Z. Zheng, Jacqueline R. Vertino, Antony D'Avirro, Benjamin Peloquin
Fluid Concepts Research
Correspondence: jeff@fluidconcepts.ai

Abstract

Reading a social situation often depends on behavior, not words alone. We introduce FriendBench, a benchmark for inferring whether two people are already familiar or are meeting as strangers, from a 20-second clip of a dyadic ice-breaker conversation. Every pair answers the same type of prompt, so only the manner of interaction can reveal the answer. Across text, audio, and video, we compare 26 models from seven companies against matched human panels over 96 balanced dyads. The best model and the human crowd are statistically indistinguishable on accuracy in every modality, but reach it differently: humans stay balanced across the two answers, while the strongest models favor "stranger." This is a difference in effective prior, not in discrimination. Richer channels help both unequally, and only humans gain from visible behavior on top of speech. We release the stimuli, human ratings, and model predictions.

1 Introduction

Social intelligence, the capacity to make sense of other people and the situations they are in, is a central component of human cognition (Adolphs, 2003; Frith and Frith, 2007). It is also an increasingly important requirement for AI systems, which now observe, mediate, and participate in human interaction in roles ranging from companion agents to meeting assistants and care monitors (Mathur et al., 2024; Sap et al., 2019). Such systems must reason about the social situation people are in, and much of that reasoning rests on multimodal behavioral cues rather than explicit verbal content, such as how people coordinate, respond, and orient toward one another (Tickle-Degnen and Rosenthal, 1990). A system that reads only the semantic content of a conversation captures just part of what social understanding requires.

We study one concrete instance of social intelligence: the capacity to recognize the relationship between two people from a brief sample of how they interact. Specifically, we ask whether an observer can tell that two people are already familiar with one another rather than meeting as strangers, without being told and without relying on what the conversation is about. This judgment is a clean probe of behavioral social perception for two reasons. First, since we have ground-truth relationship labels, its answer is a matter of fact rather than interpretation: much related work in social intelligence targets intentions, emotions, or beliefs (Premack and Woodruff, 1978; Sap et al., 2019), whose correct answer is often contestable, whereas prior familiarity has an unambiguous ground truth (i.e., two people either have met before or they have not). Second, familiarity is often unstated in a short interaction, so it must be inferred from behavior. The task therefore rules out succeeding by reading semantic content alone, and prior work shows that humans make accurate social judgments from brief samples of behavior, so-called *thin slices* (Ambady and Rosenthal, 1992).

We operationalize the task as FRIENDBENCH: binary classification, familiar versus strangers, from a 20-second clip of a dyadic ice-breaker conversation, evaluated separately across text, audio, and video modalities. Every dyad, whether familiar or strangers, responds to the same type of ice-breaker prompt, so the conversational topic is matched across the two classes and cannot by itself reveal the answer, leaving the *manner* in which the two people interact as the primary signal.

In addition to the core capability we benchmark, we examine how humans and models draw on the three modalities, and how each reaches its accuracy: by telling the two classes apart (*discrimination*) or by answering one class more often regardless of the evidence (*response bias*). Two patterns stand out. First, richer channels help both models and humans, but not equally: audio adds reliable signal over text for each, yet only humans gain a further

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 127–143
October 29, 2026 ©2026 Association for Computational Linguistics

reliable increment from the visual channel, while the strongest models are flat from audio to audiovisual. They under-exploit the visible behavior humans read. Second, humans stay close to balanced across both classes in every modality, whereas models show a larger and more idiosyncratic class bias, most clearly in the text-only condition. Both patterns concern *how* humans and models solve the task, rather than whether they succeed (Section 5).

Contributions.

- We introduce FRIENDBENCH, a multimodal benchmark for inferring dyads' prior familiarity (familiar vs. strangers) from ice-breaker clips in which every dyad answers the same type of prompt, built on the Seamless Interaction dataset (Agrawal et al., 2025).[1]
- We release a matched human-rater dataset for the task, covering 96 dyads in text, audio, and video, with roughly 90 raters per modality.
- We evaluate 26 models from seven companies (OpenAI, Google, Anthropic, Alibaba, Mistral, Thinking Machines, and Meta), spanning proprietary and open-weight systems, across all three modalities.
- We evaluate humans and models on the same stimuli under matched conditions. On accuracy, the best model and the human crowd are statistically indistinguishable in every modality. But equal accuracy is not human-like perception. The strongest models reach it by favoring "stranger," while human raters stay balanced. Using signal detection theory, we characterize this as a difference in effective prior, not discrimination (Section 5).
- We find a modality ordering both rater types obey only in part: text carries little signal for either and audio adds reliable discrimination, but the audiovisual channel gives *humans* a further reliable gain while leaving the strongest models flat. Current models capture the vocal signal but under-exploit the visible behavior human observers read (Section 5.3).

2 Related Work

Thin-slice perception of familiarity. Human observers form accurate judgments about people and relationships from very brief behavioral exposure. Ambady and Rosenthal's (1992) meta-analysis found that judgments from observations under five minutes (and often under 30 seconds) predicted objective outcomes at $r \approx .39$, with longer exposure adding little across studies.

Familiarity in particular is legible in thin slices, with friendship the best-studied case: observers tell friends from strangers from silent video (Latif et al., 2014) or brief audio (Bryant et al., 2020), and cues such as inter-turn timing distinguish them (Templeton et al., 2023). The closest precedent for our own task is the Interpersonal Perception Task (Costanzo and Archer, 1989), which asks viewers to judge from a brief unscripted clip whether two people are strangers, friends, or relatives, and scores them against an objective key. Listeners manage this from the voice alone: Montepare and Vega (1988) found that people hearing one side of a phone call told an intimate male friend from a casual one at above-chance rates, and that the cue was paralinguistic rather than lexical. Dunbar et al. (2022) show that relationship quality remains inferable from speech even after its lexical content is digitally removed. The relational signal need not come from what is said.

This motivates both our task and our 20-second window. Work that instead varies exposure within a single design does find modest gains from longer clips (Carney et al., 2007). The two results fit together if a short slice already carries most of the available signal. Our 20 seconds is therefore a conservative choice rather than an accuracy-maximizing one.

How familiarity shapes interaction. A separate literature shows where this signal comes from. Speakers tune greeting prosody to the person they address, and the form displays their stance toward the relationship (Pillet-Shore, 2012). Friends and strangers also coordinate differently. Friends produce more antiphonal laughter, meaning laughter that overlaps or immediately follows a partner's (Smoski and Bachorowski, 2003), and faster between-turn responses signal connection clearly enough for third parties to hear (Templeton et al., 2022). Even function-word matching tracks relational outcomes (Ireland et al., 2011). Familiarity also marks the transcript, since speakers use richer vocabulary and more complex grammar with familiar partners than with strangers (Luo et al., 2021). These cues ride on timing, voice, and word-

[1]The benchmark stimuli, human ratings, and model predictions are openly available at https://huggingface.co/datasets/fluid-concepts/friend-bench.

ing rather than topic, which our design isolates.

Recognizing relationships from behavior. Machine analysis of nonverbal behavior for social inference has a long history (Vinciarelli et al., 2009). One line of work predicts relationship *type* from images or video via supervised classification: PISC (Li et al., 2017) labels images as intimate, non-intimate, or no-relation, PIPA (Sun et al., 2017) annotates sixteen fine-grained relations in photo albums, and more recent work classifies asymmetric relations from the temporal dynamics of a live interaction (Tang et al., 2026).

A parallel line infers relationships from the *semantic content* of dialogue: acoustic-lexical classifiers over phone calls (Katerenchuk et al., 2014) and language models over movie-script dialogue, from relation-classification datasets such as DDRel (Jia et al., 2021) to LLM evaluations where GPT-4o infers speaker relationships well above chance (Kim et al., 2026). Both lines differ from our task in two ways: they infer relationship type rather than the presence or absence of prior familiarity, and they lean on appearance, scene, or semantic content (especially for scripted dialogue). We instead fix the conversational prompt, controlling the semantic content these methods lean on.

Two studies come closer to our question, addressing familiarity itself rather than relationship type. Maynard and Zimmerman (1984) compared acquainted and unacquainted student dyads and found that unacquainted pairs build topics through pre-topical sequences that invite category self-identification, such as year in school or major, where acquainted pairs draw on shared knowledge instead. Zhao et al. (2016) found that the temporal patterns predicting rapport in 30-second slices differ between friends and strangers.

Social reasoning in multimodal models. Multimodal models are increasingly tested for social understanding: Social-IQ (Zadeh et al., 2019) poses questions about social videos, while SIV-Bench (Kong et al., 2026) and PIVOTSBench (Zhang et al., 2026) probe reasoning about social scenes and fine-grained relations. HumanSense (Qin et al., 2026) is the closest to our setting. One of its subtasks asks a model to judge how well two people in a video know each other. Three things set our benchmark apart from these: (1) every dyad in our set answers the same type of prompt, so the topic itself cannot give the answer away; (2) our label is objective, recording whether a pair had actually met before rather than how close a viewer judges them; and (3) we gather matched human ratings in text, audio, and video for direct comparison.

Our stimuli come from the Seamless Interaction corpus (Agrawal et al., 2025). Other dyadic corpora record acquaintance too, but treat it as metadata rather than as a label for prediction. UDIVA (Palmero et al., 2021) labels each pair known or unknown, and NoXi (Cafaro et al., 2017) rates how well partners know each other.

Comparison against a human panel is becoming standard practice. Strachan et al. (2024) tested models and 1,907 people on classic theory-of-mind items, such as false belief and faux pas recognition, and traced an apparent model advantage on one of them to a response bias rather than to inference. Those items are written vignettes about mental states rather than perceptual judgments about a recorded interaction, so we take the design from that work rather than the task. Reported theory-of-mind success can also be brittle to small adversarial changes (Shapira et al., 2024), which is one reason to test naturalistic stimuli. Such biases are well documented. A model's forced choice partly reflects a prior over answer labels separable from its ability to discriminate (Zhao et al., 2021), and models can carry a systematic default answer that runs opposite to the human one (Braun, 2025). We therefore report per-class recall and signal-detection statistics rather than accuracy alone.

3 Methods

3.1 Benchmark Construction

Samples are drawn from the Seamless Interaction dataset (Agrawal et al., 2025), restricted to naturalistic interactions from three recording sites (the dataset's *vendor* field), each contributing a comparable share of dyads. We exclude the dataset's improvised interactions, in which participants are assigned a relationship and asked to act it out, since our aim is to measure this task on real, unscripted behavior rather than acted behavior. Because the benchmark is used only for zero-shot evaluation and never for training, we pool dyads across all of the dataset's predefined splits rather than restricting to one, maximizing the pool from which our quality filters and stratified design (below) can draw.

Seamless Interaction sessions include many different interaction types. We use only the Either-Or (EO) ice-breaker, a short task in which one participant poses a forced-choice hypothetical question

(e.g., "would you rather have the ability to fly or be invisible?") and the dyad discusses their answer.

We chose EO for three reasons: it standardizes the conversational context, since familiar and stranger pairs do the same thing and any signal must come from how a dyad responds rather than what it discusses; it is always the first interaction in a session, so every dyad is sampled from the same point in their interaction history; and its content carries little direct information about relationship status, so the task cannot be solved from semantic content alone. In free conversation, by contrast, status can leak through content: shared history and inside knowledge for familiars, getting-acquainted basics for strangers. A shared hypothetical question largely suppresses these cues.

An EO interaction, though centered on one either-or question, typically continues for several minutes. We sample one 20-second clip per dyad from this interaction, drawn from either its early or late portion (counterbalanced across relationship and recording site), with boundaries snapped ($\pm$5s) to the nearest turn start so clips do not begin or end mid-utterance. The rendered video stimulus retains the clip's audio track, so the video condition is audiovisual, with visible behavior together with speech, whereas the audio and text conditions each isolate a single channel. Both human raters and video models therefore receive sound as well as picture in the video condition.

Each dyad's relationship is a ground-truth label carried by the source dataset, recording whether the two participants were already acquainted (*familiar*) or previously unacquainted (*strangers*). The final set fully crosses recording site, relationship, and gender composition (same-gender/mixed-gender pair), with 8 dyads per cell, for 96 dyads that are balanced by construction (48 familiar, 48 strangers), and is participant-disjoint: no individual appears in more than one dyad. Crossing rules out recording site and gender composition as confounds (Section 4). Disjointness keeps each dyad an independent observation and guards against recognition leakage: because human raters see multiple stimuli in one session, a shared individual could be recognized from an earlier item, letting that recognition serve as the cue rather than the relationship itself.[2]

Interaction- and clip-level filters (minimum turns, in-window speaker balance, audio-track synchrony, interaction length, prompt completeness; see Table A1 in Appendix A) exclude clips that cannot support the task regardless of relationship, e.g., one participant speaking for only a few seconds or desynchronized audio tracks. An additional LLM-based audit checks that each interaction's recorded prompt label matches its content.

3.2 Human Ratings

To establish baseline human performance, we collected human ratings of relationship type in a crowdsourced study. For each stimulus, raters made a forced choice: familiar or strangers. Before rating, each rater read a short instruction screen. It explained that they would read, listen to, or watch (depending on modality) a clip from a real conversation in which the two people were playing the "Either Or" ice-breaker game, then judge whether the pair had met before (*familiar*) or were meeting for the first time (*strangers*). The full instruction text is reproduced in Appendix C.

Raters were recruited via Prolific and completed the task on GORILLA. We ran three experiments, one per modality (text, audio, video), each with a panel of roughly 90 raters (94 text, 94 audio, 92 video). Prolific prescreening and quota matching ensured a gender-balanced pool of US nationals residing in the US, with no reported hearing difficulties or autism-spectrum diagnosis, and excluded anyone who had already participated in one of our other modality panels (see Appendix D).

Raters followed a planned-missing, 6-block incomplete design (Graham et al., 2006), each rating one of six overlapping blocks and seeing 32 of the 96 dyads. Individual stimuli were in turn rated by 25–35 raters.This design lets us model raters themselves, rather than averaging one rating per stimulus. We estimate each rater's own accuracy and response bias with crossed random effects. This estimate requires each rater to contribute enough trials to be more than a one-shot data point.

Each rater additionally completed the Social Information Processing (SP) subscale of the English-language Tromsø Social Intelligence Scale (TSIS-E; Silvera et al., 2001; Grieve and Mahar, 2013) and a post-task self-report of which cues (visual, vocal, verbal, interactional) they attended to, for individual-differences analysis (Appendix E).

[2]This is not a risk for our models, which are evaluated zero-shot with no training process in which to learn identity.

3.3 Model Predictions

We evaluate 26 models from seven companies across text, audio, and video (Table 2), a fully-crossed design in which every model rates every stimulus. Unlike the human panels, no incomplete-block correction is needed for the model results. Each model receives the same prompt and response parsing. The model prompt is adapted directly from the human instruction text, so that both rater types are given the same task framing, the same "Either Or" description, and the same familiar/strangers definition. The only difference is the response mechanism: a forced-choice button for humans, a JSON label with a confidence rating for models. Models and prompts are further described in Appendix B and Appendix C, respectively.

A few audio models decline to answer on a share of trials (returning no usable familiar/strangers label) rather than committing to a forced choice. Because a refusal is not a wrong answer, we report each model's accuracy over its *answered* (covered) trials and give per-model coverage alongside it (Table 2). Scoring refusals as errors would penalize a model for declining rather than for misjudging. Coverage is near-total except for three audio models (Section 4), and the top models in each modality answer every trial, so this choice does not affect the headline comparison.

4 Benchmark Results

Table 1 summarizes, per modality, the average per-rater human, the human crowd (the per-stimulus majority vote of all raters who saw that stimulus), and the single best-performing model. For each it gives accuracy and, for the class-bias analysis in Section 5.1, per-class recall and signal-detection statistics (Stanislaw and Todorov, 1999). Table 2 gives the complete per-model breakdown with Wilson 95% confidence intervals. Figure 1 depicts the same comparison, plotting every model against the average individual human rater and the majority-vote crowd in each modality.

The best model and the human crowd trade the point-estimate lead across modalities. The model edges the crowd in text (by 5.2 points) and audio (by 3.2), while the crowd leads in video (by 5.2). Video is humans' strongest modality, and the one place the crowd, not a model, comes out ahead. None of these three gaps is statistically reliable, though. On the 96 paired dyads, an exact McNemar test (Dietterich, 1998; McNemar, 1947) never

Rater	Acc.	Fam.	Str.	d'	c
Text					
Human (indiv.)	49.9	54.6	45.3	0.00	−0.12
Human (crowd)	51.0	59.6	42.6	0.05	−0.21
`gpt_text_mini`	56.2	20.8	91.7	0.54	1.06
Audio					
Human (indiv.)	56.3**	57.2	55.5	0.32	−0.02
Human (crowd)	63.5*	64.6	62.5	0.68	−0.03
`gemini_audio_pro`	66.7**	39.6	93.8	1.21	0.86
Video					
Human (indiv.)	60.9***	58.7	63.2	0.56	0.06
Human (crowd)	71.9***	70.2	74.5	1.16	0.06
`gemini_video`	66.7**	41.7	91.7	1.12	0.77

Table 1: Human raters (average individual and majority-vote crowd) vs. the best model per modality. Acc. is accuracy; Fam. and Str. per-class recall (all %); d' and c as in Table 2. The named model in each block is that modality's best-performing (top-accuracy) system, scored over answered trials; human-indiv. accuracy is the per-rater mean (per-model Wilson CIs in Table 2). */**/***: above chance at .05/.01/.001, by exact binomial vs. 50% (crowd, best model) or 95%/99%/99.9% GLMM posterior HDI excluding 50% (individual rows).

approaches significance ($p > .4$ in every modality). Text is the weakest modality for humans: at chance individually and barely above it as a crowd. But the best text model is also not above chance (56.2%, $p = .26$). Text is therefore less a model win than a modality where neither rater type finds a signal.

Only three models are individually above chance at this sample size, all from Google: `gemini_audio_pro` and `gemini_audio` in audio, and `gemini_video` in video. No text model clears chance. Apart from the two top-accuracy systems, the best remaining model in each modality reaches only 56.2% (text, `gpt_text_mini`), 63.5% (audio, `gemini_audio`), and 58.3% (video, `gemini_video_thinking`). Most models cluster near chance on raw accuracy. The per-class recall and criterion columns already hint at the analysis in Section 5.1: several models carry very large response biases regardless of accuracy. In audio the bias is large but inconsistent in direction. `voxtral_audio` labels every dyad "familiar" (100%/0%, $c = -2.32$), while `qwen_audio` does the opposite (0%/100%, $c = 2.19$). In text most models strongly favor "stranger," with `muse_spark_text_high` recovering only 8.3% of familiar pairs ($c = 1.40$). A near-chance accuracy can thus conceal a strongly one-sided response pattern rather than balanced guessing.

Three audio models also decline a substantial

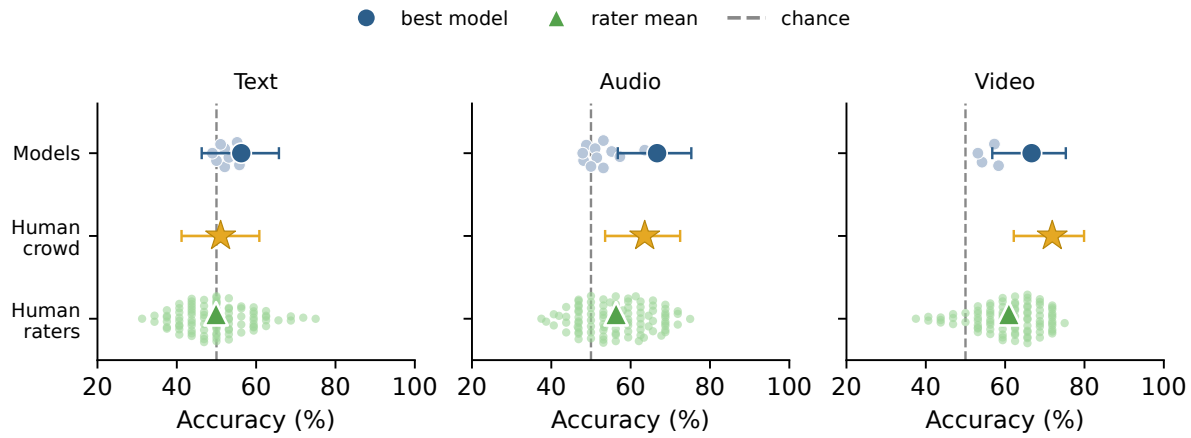


Figure 1: Human vs. model accuracy by modality. Each rater point covers only the ∼32 trials seen by that rater, so the rater cloud is widened by sampling noise and its upper tail is not a band of expert raters (Section 5.4). Horizontal bars are Wilson 95% CIs for the best model and crowd, not tests of human–model or cross-modality differences (Section 4, Section 5.3).

share of trials: `gpt_audio` answers only 45%, `gpt_audio_mini` 28%, and `qwen_audio` 71%. They sit at chance on the trials they do answer, so their behavior is better summarized as declining-plus-guessing.

5 Performance Analysis

5.1 Class Bias: Humans vs. Models

Humans and the best-performing models are close to parity on accuracy (Section 4), but do they reach it the same way? To separate *discrimination* from *response bias*, we read the per-class recall and signal-detection columns of Table 1, treating "familiar" as the signal class: d' measures how well a rater tells the classes apart, and the criterion c how far its decision rule favors one answer ($c > 0$ toward "stranger," $c < 0$ toward "familiar"). Two patterns stand out. Humans stay near-balanced: the criterion is within 0.23 of zero in every modality, and per-class recall gaps never exceed 10 points for pooled raters. The best text and audio models instead reach their accuracy largely through response bias: `gemini_audio_pro` labels 93.8% of strangers correctly but recovers only 39.6% of familiar pairs ($c = 0.86$), and the best text model, `gpt_text_mini`, is more extreme (91.7% vs. 20.8%, $c = 1.06$). Across the roster this is the rule: in text almost every model favors "stranger" by 50–85 recall points, and in audio the bias is large but inconsistent in direction (Table 2).

The gap holds even in video, where models come closest to humans. There the best model (`gemini_video`) still favors "stranger" ($c = 0.77$; 91.7% of strangers but only 41.7% of familiar pairs), while the human crowd stays balanced ($c = 0.06$) at slightly higher discrimination ($d' = 1.16$ vs. 1.12). The two reach comparable d', but only the crowd does so without skewing its decision rule. Models thus fail to recreate the human balanced-recall pattern in *any* modality, including their strongest: several of the highest-accuracy systems (Table 2) buy that accuracy with a skewed criterion rather than sharper discrimination.

5.2 Wisdom of the Crowd

Relative to the average single rater, pooling human raters via majority vote barely moves accuracy in text (49.9%→51.0%, +1.1 points) but raises it substantially in audio (56.3%→63.5%, +7.2 points) and video (60.9%→71.9%, +11.0 points). In each pair the first number is the mean single-rater accuracy and the second the full-panel majority vote. Figure 2 shows the full crowd-growth curves: majority-vote accuracy rises with crowd size k in audio and video but stays flat near chance in text. The audio and video gains therefore reflect real signal, not an artifact of pooling.

5.3 Modality Comparison

Text is the weakest modality for both humans and models, and richer channels help both, though not to the same degree, as we show below. To test the modality ordering while accounting for the repeated-measures structure of the human data, we fit a Bayesian crossed random-effects logistic

Model	Acc.	95% CI	Cov.	Fam.	Str.	d'	c
Text							
gpt_text_mini	56.2	[46.3, 65.7]	100	20.8	91.7	0.54	1.06
gemini_text_thinking	55.8	[45.8, 65.4]	99	27.7	83.3	0.36	0.76
gemini_text	55.2	[45.3, 64.8]	100	25.0	85.4	0.36	0.84
gpt_text	53.1	[43.2, 62.8]	100	29.2	77.1	0.19	0.63
inkling_text	52.1	[42.2, 61.8]	100	16.7	87.5	0.17	1.03
claude_text	52.1	[42.2, 61.8]	100	14.6	89.6	0.19	1.12
muse_spark_text_high	51.0	[41.2, 60.8]	100	8.3	93.8	0.14	1.40
mistral_text	50.0	[40.2, 59.8]	100	87.5	12.5	0.00	−1.11
muse_spark_text	49.0	[39.2, 58.8]	100	10.4	87.5	−0.10	1.16
Audio							
gemini_audio_pro	66.7**	[56.8, 75.3]	100	39.6	93.8	1.21	0.86
gemini_audio	63.5*	[53.6, 72.5]	100	81.2	45.8	0.76	−0.48
muse_spark_audio	57.3	[47.3, 66.7]	100	20.8	93.8	0.67	1.13
gemini_audio_thinking	55.2	[45.3, 64.8]	100	64.6	45.8	0.26	−0.23
muse_spark_audio_high	53.1	[43.2, 62.8]	100	18.8	87.5	0.25	0.99
gpt_audio_1_5	53.1	[43.2, 62.8]	100	87.5	18.8	0.25	−0.99
qwen_audio	51.5	[39.8, 62.9]	71	0.0	100.0	0.02	2.19
inkling_audio	51.0	[41.2, 60.8]	100	41.7	60.4	0.05	0.23
voxtral_audio	50.0	[40.2, 59.8]	100	100.0	0.0	0.00	−2.32
gpt_audio	48.8	[34.6, 63.2]	45	100.0	4.3	0.45	−1.76
gpt_audio_mini	48.1	[30.7, 66.0]	28	25.0	81.8	0.18	0.72
qwen_omni	47.9	[38.2, 57.8]	100	16.7	79.2	−0.15	0.87
Video							
gemini_video	66.7**	[56.8, 75.3]	100	41.7	91.7	1.12	0.77
gemini_video_thinking	58.3	[48.3, 67.7]	100	27.1	89.6	0.62	0.91
gemini_video_pro	57.3	[47.3, 66.7]	100	18.8	95.8	0.77	1.25
muse_spark_video	54.2	[44.2, 63.8]	100	14.6	93.8	0.44	1.24
muse_spark_video_high	53.1	[43.2, 62.8]	100	10.4	95.8	0.42	1.42

Table 2: Per-model results, sorted within modality. Acc. is accuracy (%) over answered trials and CI is its Wilson 95% confidence interval; Cov. is coverage (% answered rather than declined). Fam. and Str. are per-class recall (%): share of familiar and of stranger dyads correctly labeled. d' and criterion c treat "familiar" as the signal class (log-linear corrected; Hautus, 1995), so $c > 0$ indicates a bias toward "stranger" and $c < 0$ toward "familiar." **/*: accuracy above chance at $p < .01$/$p < .05$ (exact binomial vs. 50%).

model (Gelman et al., 2014) to per-trial human correctness (Bernoulli likelihood, logit link), with a fixed effect of modality and crossed random intercepts for rater and for dyad (Section 3.2). This is the design the block structure was built to support. We estimate it with the `bambi` interface to PyMC (Capretto et al., 2022), using default weakly-informative, data-scaled priors (wide normal priors on the fixed effects and half-normal hyperpriors on the random-effect standard deviations), and sample with NUTS (4 chains, 1,000 warmup and 1,000 post-warmup draws each; all $\hat{R} \approx 1.00$). We report population-averaged (marginal) accuracies and contrasts as posterior means with 95% highest-density intervals (HDIs, Makowski et al., 2019).

The marginal accuracies confirm the raw pattern: text 50.0% (95% HDI $[46.7, 53.6]$), audio 56.6% $[52.2, 60.6]$, video 60.9% $[56.9, 64.9]$. The pairwise contrasts are decisive: audio exceeds text by 6.5 points (HDI $[4.1, 8.9]$), video exceeds audio by 4.5 points ($[1.9, 6.8]$), and video exceeds text by 10.9 points ($[8.5, 13.3]$), each with posterior $P > 0$ of at least 0.999. Text alone is statistically indistinguishable from chance (posterior probability of exceeding 50% only 0.50). Because the video condition is audiovisual (Section 3.1), its edge over audio reflects the *added* value of visible behavior on top of speech, not vision in isolation. Video is best read as an audiovisual upper bound rather than a vision-only channel.

The model side matches the human floor but not the human ceiling. The best model climbs from text (56.2%) to audio (66.7%) but gains nothing from the audiovisual channel (66.7%), and among models evaluated in both audio and video only Gemini Flash improves (63.5% → 66.7%) while Gemini Pro (66.7% → 57.3%) and Muse Spark (57.3% → 54.2%) decline. (Cross-modality model means rise monotonically: 52.7, 53.9, 57.9% for text, audio, video. But the video roster is small and skewed toward stronger companies, so we read the best-model and within-model trajectories rather

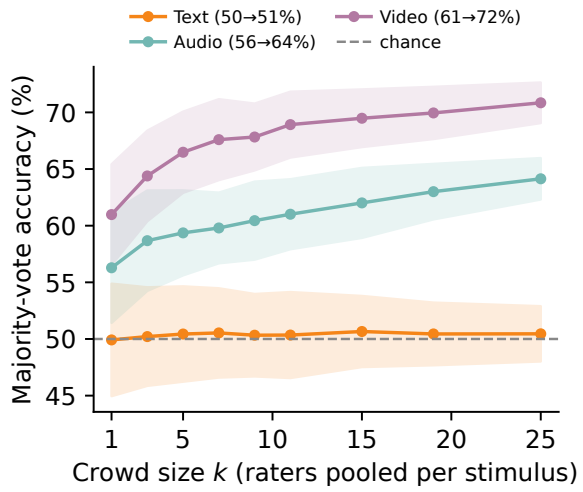


Figure 2: Wisdom of the crowd. Majority-vote accuracy as a function of human crowd size k, with raters sampled per stimulus from that stimulus's own rater panel (mean $\pm$ SD over 300 bootstrap draws). Legend gives single-rater $\rightarrow$ full-panel crowd accuracy per modality.

than the pooled mean.) Humans reliably convert visible behavior into accuracy on top of speech, but the strongest models do not. They capture the vocal signal but under-exploit the visual channel. That channel is where a benchmark of *multimodal* social perception most expects a model to gain.

That text is the hardest modality is consistent with the benchmark's design rather than a defect of it. The EO ice-breaker was chosen precisely so that transcript content carries little direct information about relationship status (Section 3.1). The weak transcript performance of both humans and models is the expected consequence of that choice. It also helps explain why response bias is most visible in text: with little signal available, a rater's responses reflect its bias more than the stimulus.

5.4 Rater Individual Differences

Because each rater contributes many trials by design (Section 3.2), we can estimate individual accuracy and relate it to rater traits. Trait social intelligence (SP subscale; reliable in every panel, $\alpha = 0.88$–0.92) is essentially uncorrelated with accuracy in the two modalities where the task is doable: audio $r = 0.03$ ($p = .78$) and video $r = -0.12$ ($p = .24$). The only significant association is in text ($r = 0.29$, $p = .004$), the modality where average accuracy is at chance, so we read it with caution. Self-reported cue use is similarly flat: raters most often report attending to interactional cues (rapport, responsiveness), but cue-use scores rarely predict accuracy (Appendix E).

The individual-rater cloud in Figure 1 is correspondingly wide, with the best raters near 75% in every modality. Its spread should not be read as a stable band of expert observers, though. Each rater's accuracy comes from only the $\sim$32 trials they saw, so the upper tail is close to what sampling noise alone would produce. The best model therefore sits within the human distribution, not above it, even as almost no individual beats the aggregated crowd (1 of 92 in video). The crossed random-effects model makes the point directly: the dyad random-intercept SD (0.67–0.95 across modalities, logit scale) is six to seven times the rater SD (0.10–0.14). Who the rater is thus matters far less than which dyad they were rating.

5.5 Dyad Difficulty

Treating the per-dyad random intercepts from the crossed random-effects model as Rasch-style item-easiness parameters (Rasch, 1960; de Boeck and Wilson, 2004), we find that difficulty is largely a property of the conversation, not of the modality through which it is observed. Estimated dyad easiness correlates positively across all modality pairs (text–audio $r = 0.52$, text–video $r = 0.33$, audio–video $r = 0.61$; all $p < .01$): a dyad that is hard to read from one modality tends to be hard from the others as well. A handful of dyads sit below chance in every modality, acting as systematic "lures" that most raters misread alike (Appendix F).

Humans and models tend to find the same dyads hard, though how strongly depends on how it is measured. Correlating per-dyad human-crowd accuracy with per-dyad accuracy pooled over all models yields $r = 0.56$ (audio) and $r = 0.34$ (video), both significant ($p < .01$; Appendix Figure A1), but only $r = 0.20$ in text ($p = .046$). This raw correlation understates the shared difficulty. It blends two things: how hard a dyad is to read, and which class it belongs to. Humans stay balanced while the models favor "stranger" (Section 5.1), so the two rater types tend to miss different classes. Isolating difficulty from this class split, by correlating within each true class, brings the agreement out clearly: $r = 0.61$ (text), 0.58 (audio), and 0.44 (video), all $p < .001$. So a substantial part of what makes a pair legible or illegible is a property of the interaction shared across observers, even though the two reach their answers by different rules.

6 Discussion

For a benchmark of multimodal social perception, the clearest result is about the modalities themselves. Text carries little relational signal for anyone, by design: the shared ice-breaker prompt suppresses lexical content (Section 3.1). Richer channels help both humans and models, but asymmetrically. Humans improve reliably at every step, from text to audio to audiovisual (Section 5.3), gaining a credible increment from visible behavior on top of speech. The strongest models capture the audio gain but not the visual one: their top accuracy is identical in audio and audiovisual (66.7%). The sharpest human–model difference is thus not whether models can read relationships but whether they exploit the visual channel as humans do. This gap has applied stakes. The companion agents, meeting assistants, and care monitors that motivate the task must read social situations from behavior, and the visible cues humans exploit on top of speech are exactly what current models miss.

On overall accuracy, the best models have drawn level with the human crowd. In every modality the two are statistically indistinguishable. Point estimates put the model slightly ahead in text and audio and the crowd slightly ahead in video (Table 1), but no gap survives a paired McNemar test ($p > .4$ throughout; Section 4).

It may seem surprising that human accuracy is not higher, since people are known to read relationships well from brief exposure (Section 2). That prior work sets a realistic bar rather than a high one. Its accuracy effects are moderate in size, near $r = .39$, and not close to ceiling. Our task is also hard by design. Each judgment rests on a single 20-second clip, the response is a forced choice, and the balanced set removes any base-rate cue that a rater could use as a shortcut. The human signal is nonetheless real. The crowd reaches about 72% in video, and the best raters approach 75% in every modality (Section 5.4).

The difference is in how humans and models use their two answers. Discrimination (d') measures how well a rater tells the classes apart. Response bias (criterion c; Table 1) measures how far it favors one answer. Human raters stay balanced across the two answers in every modality. The strongest models instead favor "stranger," giving that answer more often regardless of the pair. Video makes this clearest. There the best model tells the classes apart as well as the crowd ($d' \approx 1.1$), but it reaches that accuracy by favoring "stranger" where the crowd stays balanced.

In signal-detection terms, this is a difference in *effective prior*: the models behave as though strangers were the more common answer and demand more evidence before saying "familiar." Humans behave as though the classes were equally likely, which is true in our balanced set. This describes the response distribution, not a mechanism, and the prior is inferred from behavior rather than verified. A pure criterion shift is correctable: recentering to the known base rate would raise accuracy with no gain in discrimination. So a biased model's raw accuracy can understate its discrimination.

We cannot say from these data why models adopt this prior. One simple explanation is that strangers are more often the correct answer in the world, so favoring that answer would be well calibrated rather than mistaken. This does not fit our results well. The size of the bias, and even its direction, varies widely from model to model (Table 2), which is not what a single shared world prior would produce. We therefore treat this skew as a property of each model's decision rule, not as evidence that models have learned the true base rate.

At the same time, humans and models are not perceiving unrelated things. Dyad difficulty is partially shared and partially transfers across modalities (Section 5.5), so part of what makes a pair legible or illegible is a property of the interaction itself, not of the observer or channel. The dissociation is therefore specific: the two agree on *which* dyads are hard but diverge on the *decision rule* they apply. Methodologically, these results argue for evaluating social-perception models with more than a single accuracy number: per-class recall, signal-detection statistics, and item-level difficulty each revealed structure that accuracy alone obscured, and are cheap to report once the underlying predictions are available.

Overall, the strongest models now match the human crowd on accuracy, but not on how they reach it. They find the same conversations hard. Yet they systematically favor "stranger," where human raters stay balanced, and they do not convert the visual channel into accuracy. This benchmark is built to track whether models can be brought to read relationships as humans do, balanced across classes and drawing on visible behavior rather than speech alone.

Limitations

This paper reports only the ice-breaker (first-interaction) task from the Seamless Interaction dataset. Findings about modality strength and class bias may not generalize to conversations later in a relationship or session, or to different prompt structures. All familiar subtypes (friends, family, romantic partner, coworkers, familiar_other) are collapsed into a single "familiar" class, because the balanced design does not have enough dyads per subtype to power a multi-class analysis. A six-class relationship-type task is defined in the broader project but is out of scope here, so any within-familiar heterogeneity is invisible to this binary framing.

Only two-person interactions are studied, and relationship inference in larger groups may draw on cues not captured here. The balanced evaluation set is also modest in size (96 dyads, one clip each), so per-model accuracy intervals are wide and only the strongest models clear chance individually. The class-bias and difficulty patterns we emphasize are more robust than any single model's rank.

Our human baseline is drawn entirely from US-resident raters (Section 3.2). Relationship-perception cues can be culturally specific, so this baseline may not represent human performance in other populations. Finally, even the "naturalistic" subset was recorded in a fixed-camera motion-capture studio, with wired lapel microphones and a posed ice-breaker prompt. This is not truly in-the-wild interaction, so findings may not transfer to less controlled contexts such as phone video or casual settings.

We also do not identify which cues drive these judgments. Our measures show whether each rater type separates the two classes and how it biases its answers, but not what in the clip carries the signal, whether that is prosody, word choice, facial or bodily movement, or the timing of the exchange. We do collect rater self-reports of the cues people say they attended to (Section 5.4), but these rarely predict accuracy and do not tell us which cues actually matter, for humans or for models. Identifying them, for example through controlled removal of parts of the input or through probing of model representations, is useful future work.

Our evaluation set is also balanced 50/50 between familiar and stranger dyads by construction, which makes accuracy a clean measure of discrimination but base-rate-specific. A reader might ask whether the models' skew toward "stranger" is not miscalibration but a well-calibrated prior for a world in which two people recorded together are more often strangers, penalized only by our artificial balance. This does not threaten our central claim. That claim is the *difference* in criterion between humans and models on matched stimuli. Neither rater type was told the base rate, so this difference does not depend on the true base rate, and neither does d', which is base-rate-invariant. It does mean we measure discrimination, not deployment calibration. We do not read the balanced accuracy as a deployment estimate. As noted in Section 6, this bias is idiosyncratic across models, which argues against calibration to any single real-world base rate. Measuring calibration against realistic base rates would complement the discrimination-focused evaluation we report here.

Each model is also evaluated under a single prompt, adapted from the human instructions (Section 3.3). Because model responses can be sensitive to prompt wording, the exact biases in Table 2 may shift under other phrasings. A prompt-robustness sweep is worthwhile future work. The human–model comparison holds the task framing fixed for both rater types, so it is less exposed to this concern than any single model's bias estimate read on its own.

Ethics Statement

Informed consent and compensation. Rater recruitment and prescreening are described in Section 3.2 and Appendix D. Before viewing any stimulus, each rater read a description of the study and affirmatively agreed to three statements: that they had read and understood the information, that they were free to withdraw at any time, by closing the browser tab, without giving a reason, and that they agreed to take part. Participation was voluntary and could be ended at any point without penalty. Sessions took roughly 15–30 minutes depending on modality, and raters were compensated at an effective rate of approximately $10–12 per hour, above Prolific's fair-pay guidance. The task was minimal-risk: raters viewed short, benign clips of consenting conversation partners and answered non-sensitive perceptual and self-report questions.

Privacy and data minimization. We collected no personally identifying information from raters. Beyond the task responses and self-report questionnaires analyzed here, we recorded only the coarse

attributes used for prescreening and quota matching (Section 3.2), keyed to a pseudonymous Prolific identifier we do not link to any real-world identity. All released rater data are de-identified.

Institutional review. This research was conducted outside a university setting and was not reviewed by an institutional review board. Because it collected no personally identifying information from raters and involved only minimal-risk procedures, it did not fall under IRB oversight. We nonetheless followed standard human-subjects safeguards: voluntary informed consent, the right to withdraw without penalty, fair compensation, minimal-risk stimuli, and data de-identification.

Stimulus source. All interaction clips are drawn from the publicly released Seamless Interaction dataset (Agrawal et al., 2025), whose participants consented to the recording and research use of their audio and video. We use these data in accordance with the dataset's CC-BY-NC 4.0 license (attribution, non-commercial use only): the rendered benchmark clips are redistributed under the same CC-BY-NC 4.0 terms, with attribution to the Seamless Interaction dataset. We do not attempt to re-identify or contact any recorded individual.

Intended use and risks. The benchmark is intended for evaluating and auditing the social-perceptual behavior of models, including the response biases we document (Section 5.1). Inferring familiarity from behavior could in principle support surveillance or profiling. We release the benchmark to enable research on and scrutiny of such capabilities, not their deployment. Given the modest accuracy and pronounced, idiosyncratic biases we observe (Section 4), we caution against using these models or this task to make consequential judgments about real individuals.

References

Ralph Adolphs. 2003. Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3):165–178.

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, and 65 others. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset.

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274.

Daniel Braun. 2025. Acquiescence bias in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11341–11355. ACL.

Gregory A. Bryant, Christine S. Wang, and Riccardo Fusaroli. 2020. Recognizing affiliation in colaughter and cospeech. *Royal Society Open Science*, 7(10):201092.

Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: Multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, pages 350–359, New York, NY, USA. Association for Computing Machinery.

Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A. Martin. 2022. Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103:1–29.

Dana R. Carney, C. Randall Colvin, and Judith A. Hall. 2007. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072.

Mark Costanzo and Dane Archer. 1989. Interperting the expressive behavior of others: The interpersonal perception task. *Journal of Nonverbal Behavior*, 13(4):225–245.

Paul de Boeck and Mark Wilson, editors. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer Science & Business Media.

Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923.

R. I. M. Dunbar, Juan-Pablo Robledo, Ignacio Tamarit, Ian Cross, and Emma Smith. 2022. Nonverbal Auditory Cues Allow Relationship Quality to be Inferred During Conversations. *Journal of Nonverbal Behavior*, 46(1):1–18.

Chris D. Frith and Uta Frith. 2007. Social Cognition in Humans. *Current Biology*, 17(16):R724–R732.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2014. *Bayesian Data Analysis*, 3rd edition. CRC Press, Boca Raton, FL.

John W Graham, Bonnie J Taylor, Allison E Olchowski, and Patricio E Cumsille. 2006. Planned missing data designs in psychological research. *Psychological Methods*, 11(4):323–343.

Rachel Grieve and Doug Mahar. 2013. Can social intelligence be measured? Psychometric properties of the Tromsø Social Intelligence Scale – English Version. *The Irish Journal of Psychology*, 34(1):1–12.

Kilem L Gwet. 2021. *Handbook of Inter-Rater Reliability: Chance-corrected Agreement Coefficients*, 5th edition, volume 1. AgreeStat Analytics.

Michael J. Hautus. 1995. Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, 27(1):46–51.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Qi Jia, Hongru Huang, and Kenny Q. Zhu. 2021. DDRel: A New Dataset for Interpersonal Relation Classification in Dyadic Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13125–13133.

Denys Katerenchuk, David Guy Brizan, and Andrew Rosenberg. 2014. "was that your mother on the phone?": Classifying interpersonal relationships between dialog participants with lexical and acoustic properties. In *Proc. Interspeech 2014*, pages 1831–1835.

Eunsu Kim, Junyeong Park, Juhyun Oh, Kiwoong Park, Seyoung Song, A. Seza Doğruöz, Alice Oh, and Najoung Kim. 2026. Are they lovers or friends? Evaluating LLMs' Social Reasoning in English and Korean Dialogues. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23431–23451, San Diego, California, United States. Association for Computational Linguistics.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2026. SIV-Bench: A Video Benchmark for Social Interaction Understanding and Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37379–37403, San Diego, California, United States. Association for Computational Linguistics.

Nida Latif, Adriano V. Barbosa, Eric Vatikiotis-Bateson, Monica S. Castelhano, and K. G. Munhall. 2014. Movement Coordination during Conversation. *PLOS ONE*, 9(8):e105036.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2017. Dual-Glance Model for Deciphering Social Relationships. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2659.

Minxia Luo, Rudolf Debelak, Gerold Schneider, Mike Martin, and Burcu Demiray. 2021. With a little help from familiar interlocutors: Real-world language use in young and older adults. *Aging & Mental Health*, 25(12):2310–2319.

Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, 10.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing Social Intelligence in AI Agents: Technical Challenges and Open Questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, Miami, Florida, USA. Association for Computational Linguistics.

Douglas W. Maynard and Don H. Zimmerman. 1984. Topical talk, ritual and the social organization of relationships. *Social Psychology Quarterly*, 47(4):301–316.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157.

Joann M. Montepare and Cynthia Vega. 1988. Women's vocal reactions to intimate and casual male friends. *Personality and Social Psychology Bulletin*, 14(1):103–113.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapes, Alexa Mosegui, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. 2021. Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–12.

Danielle Pillet-Shore. 2012. Greeting: Displaying stance through prosodic recipient design. *Research on Language and Social Interaction*, 45(4):375–398.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Zheng Qin, Ruobing Zheng, Yabing Wang, Tianqi Li, Yi Yuan, Jingdong Chen, and Le Wang. 2026. HumanSense: From Multimodal Perception to Empathetic Context-Aware Responses Through Reasoning MLLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):24973–24981.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 5835 S.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

David Silvera, Monica Martinussen, and Tove I. Dahl. 2001. The Tromsø Social Intelligence Scale, a self-report measure of social intelligence. *Scandinavian Journal of Psychology*, 42(4):313–319.

Moria Smoski and Jo-Anne Bachorowski. 2003. Antiphonal laughter between friends and strangers. *Cognition and Emotion*, 17(2):327–340.

Harold Stanislaw and Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A Domain Based Approach to Social Relation Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3481–3490.

Wang Tang, Fethiye Irmak Dogan, Linbo Qing, and Hatice Gunes. 2026. AsyReC: A Multimodal Graph-Based Framework for Spatio-Temporal Asymmetric Dyadic Relationship Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3):3693–3708.

Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont, and Thalia Wheatley. 2023. Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1875):20210471.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Shuxiang Zhang, Yiting Yin, Wenxuan Song, Yuhang Wu, and Miao Liu. 2026. PIVOTSBench: Evaluating Fine-Grained Interpersonal Relationship Reasoning in Multimodal Large Language Models. *Preprint*, arXiv:2606.23092.

Ran Zhao, Tanmay Sinha, Alan W. Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Intelligent Virtual Agents*, pages 218–233, Cham. Springer International Publishing.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Benchmark Quality Filters

Every candidate interaction and clip must pass the automatic quality filters in Table A1 before entering the stratified selection pool (Section 3.1). Interaction-level filters (F1, F4, F5) gate the whole source interaction; clip-level filters (F2, F3) gate the specific 20-second window sampled from it. Clips are 20 s long with boundaries snapped within ± 5 s to the nearest turn start. These filters remove clips that cannot support the task regardless of relationship; they are followed by the LLM-based prompt-adherence audit (Section 3.1), which verifies that each interaction's recorded prompt label matches its actual content.

B Model Inference Configuration

Table A2 lists the inference configuration for each of the 26 models. All models received the same prompt construction and response parsing (Section 3.3); the settings below cover per-model decoding and, where applicable, reasoning parameters. Cloud models were queried at temperature 0 with a fixed seed (42) wherever the API exposed them; the three locally-run open-weight models (Qwen2-Audio, Qwen2.5-Omni, Voxtral) used greedy decoding (`do_sample=False`), which is deterministic without a seed. Answer length was capped at 150–300 tokens for non-reasoning models, while reasoning models received a larger combined answer-plus-reasoning budget (8,000 tokens for the Gemini thinking and pro variants; 4,000 for Inkling and Muse Spark) to avoid truncating the response.

C Task Instructions and Prompts

Both human raters and models were given the same task framing, "Either Or" game description, and familiar/strangers definition, differing only in how a response was collected. Human raters selected an answer with a button; models were additionally instructed to emit a JSON object with a label and confidence. Modality-specific slots (shown here in brackets, Text/Audio/Video order) were filled per panel or per clip.

Human rater instructions.

```
You will [read / listen to / watch] a series
of [transcripts / audio clips / video clips]
taken from real conversations between two
people.

These people are playing a game called
"Either Or," where they discuss whether they
would prefer one thing -- for example, the
ability to fly -- or an alternative, such as
the ability to breathe underwater.

Your job is to pay close attention to each
conversation and determine whether the two
people have met before -- meaning they are
familiar, such as friends, family members,
coworkers, or romantic partners -- or are
meeting for the first time in this
interaction -- meaning they are strangers.
```

Raters then made a forced choice: *familiar* or *strangers*.

Model prompt.

```
[Read / Listen to / Watch] the following
[transcript / audio clip / video clip], taken
from a real conversation between two people.

These people are playing a game called
"Either Or," where they discuss whether they
would prefer one thing -- for example, the
ability to fly -- or an alternative, such as
the ability to breathe underwater.

Your job is to pay close attention to the
conversation and determine whether the two
people have met before -- meaning they are
familiar, such as friends, family members,
coworkers, or romantic partners -- or are
meeting for the first time in this
interaction -- meaning they are strangers.

For each clip, return:
  {
    "relationship_label": str {FAMILIAR|STRANGER},
    "confidence": float [0, 1],
    "reason": str
  }

Use the following confidence scale:
  0 = Not at all confident, ...,
  0.5 = Moderately confident, ...,
  1 = Extremely confident

Do not provide extra text outside the JSON
object.
```

D Human Rating Study Details

Raters were recruited on Prolific with platform-level quality screening in addition to the demographic criteria in Section 3.2: a Prolific approval rate of 99–100% and at least 10 prior submissions. Under the planned-missing, 6-block design (Graham et al., 2006), each rater was assigned to a single block and each of the 96 dyads appeared in two of the six blocks, so every dyad was rated by an overlapping subset of raters (25–35 per dyad). By design each rater judged 32 dyads; in the audio panel a platform issue left 13 of the 94 raters with 29–31 completed trials, while all text and video raters completed the full 32.

ID	Level	Criterion
F1	Interaction	Both recorded EO prompts are free of placeholder/boilerplate text (e.g., "you will each see the same prompt," "given different lists of sentences") that signals the true question was not logged.
F2	Clip	≥ 4 merged conversational turns within the clip window.
F3	Clip	Each speaker contributes ≥ 3 s of speech within the clip window.
F4	Interaction	$< 35\%$ of cross-track turn-start pairs coincide within 0.2 s, rejecting duplicated or desynchronized audio tracks.
F5	Interaction	≥ 90 s of speech in the interaction.

Table A1: Automatic quality filters applied during clip generation and the EO interaction audit. Interaction-level filters gate the source interaction; clip-level filters gate the sampled 20 s window. Thresholds are the generator/auditor constants (`scripts/generate_samples_rapport.py`, `scripts/audit_eo_interactions.py`).

Inter-rater agreement. Agreement among human raters is slight, as expected for a difficult perceptual task with a balanced label set. Fleiss's κ (computed per stimulus over whichever raters saw it, accommodating the incomplete-block design) is 0.08 for text and 0.17 for both audio and video. Bennett's S is essentially identical (0.09, 0.17, 0.17 for text, audio, video), as expected when the two classes are balanced (Gwet, 2021). The near-zero text agreement is consistent with individual text raters performing at chance (Table 1): raters are not converging on a shared, reliable text cue.

E Rater Individual Differences

Table A3 reports, per modality, the reliability of the Social Information Processing (SP) subscale and its correlation with per-rater accuracy. The subscale is highly reliable in every panel, but its correlation with accuracy is null in audio and video and positive only in text, the modality where accuracy is at chance. We therefore do not interpret it as evidence of a general skill advantage.

For self-reported cue use, raters most often reported attending to interactional cues (rapport, responsiveness) across all modalities, but individual cue-attendance scores rarely predicted accuracy (only a handful of per-cue correlations reached $p < .05$, with no stable cross-modality pattern). Full cue-use heatmaps are provided with the released analysis notebooks.

The per-rater accuracy clouds in Figure 1 show the full distribution of per-rater accuracy in each modality (94 text, 94 audio, 92 video raters; each rater's accuracy over the 32 dyads in their assigned block, 29–32 for a few audio raters affected by a platform issue). The clouds are wide, but much of that width is sampling noise: with only $\sim$32 trials per rater, binomial variation around a common ability already reproduces most of the observed spread, including the upper tail that appears to exceed the best model (Section 5.4). The distribution should not be read as a stable ordering of raters by skill. This is also why per-rater accuracy is best modeled with partially-pooled crossed random effects (Section 3.2), which shrink noisy individual estimates, rather than trusting raw per-rater rates or treating each rater as a one-shot observation. Text raters cluster tightly around chance, consistent with the near-zero text agreement reported in Appendix D.

F Dyad Difficulty

Table A5 lists the eight hardest and eight easiest dyads by mean model-estimated probability of a correct human response, from the crossed random-effects model of Section 5.5. Several of the hardest dyads fall below chance in every modality, functioning as systematic lures. Estimated easiness correlates across modality pairs (text–audio $r = 0.52$, text–video $r = 0.33$, audio–video $r = 0.61$; all $p < .01$), indicating that difficulty is largely a property of the conversation rather than the observation channel.

Table A4 decomposes the human–model difficulty correlation reported in Section 5.5. The raw pooled correlation (human-crowd accuracy vs. accuracy pooled over all models) is strong in audio, moderate in video, and only marginal in text. The weak raw text value is a suppression effect rather than absence of shared structure: humans and models carry opposing class biases (Section 5.1), so their per-dyad accuracies partly anti-align on the familiar/stranger axis. Partialling out the true class raises the within-class correlation to $r = 0.61$ (text), 0.58 (audio), and 0.44 (video), all $p < .001$. This within-class correlation measures fine-grained difficulty that is not reducible to a two-way class effect. The shared difficulty is, however, partly a crowd-level property: repeating the analysis with the single strongest model per modality in place of the pooled estimate leaves it robust only in video

Model	API / model ID	Temp.	Seed	Reasoning
Text				
gpt_text	gpt-4o	0	42	—
gpt_text_mini	gpt-4o-mini	0	42	—
gemini_text	gemini-3.5-flash	0	42	off
gemini_text_thinking	gemini-3.5-flash	0	42	dynamic
claude_text	claude-opus-4-8	—[a]	—[a]	off[a]
mistral_text	mistral-large-latest	0	42	—
inkling_text	thinkingmachines/Inkling	0	42[b]	on
muse_spark_text	muse-spark-1.1	—[c]	—[c]	minimal
muse_spark_text_high	muse-spark-1.1	—[c]	—[c]	high
Audio				
gpt_audio	gpt-audio	0	42	—
gpt_audio_1_5	gpt-audio-1.5	0	42	—
gpt_audio_mini	gpt-audio-mini	0	42	—
gemini_audio	gemini-3.5-flash	0	42	off
gemini_audio_pro	gemini-pro-latest	0	42	dynamic[d]
gemini_audio_thinking	gemini-3.5-flash	0	42	dynamic
qwen_audio	Qwen/Qwen2-Audio-7B-Instruct	greedy	—	—
qwen_omni	Qwen/Qwen2.5-Omni-7B	greedy	—	—
voxtral_audio	mistralai/Voxtral-Mini-3B-2507	greedy	—	—
inkling_audio	thinkingmachines/Inkling	0	42[b]	on
muse_spark_audio	muse-spark-1.1	—[c]	—[c]	minimal
muse_spark_audio_high	muse-spark-1.1	—[c]	—[c]	high
Video				
gemini_video	gemini-3.5-flash	0	42	off
gemini_video_pro	gemini-pro-latest	0	42	dynamic[d]
gemini_video_thinking	gemini-3.5-flash	0	42	dynamic
muse_spark_video	muse-spark-1.1	—[c]	—[c]	minimal
muse_spark_video_high	muse-spark-1.1	—[c]	—[c]	high

Table A2: Per-model inference configuration (26 models, seven companies; roster matches Table 2). *Temp.* and *Seed* are the decoding temperature and random seed passed to the API; `greedy` marks the locally-run open-weight models decoded with `do_sample=False` (deterministic, no seed). *Reasoning*: "—" = non-reasoning model; "off" = reasoning-capable but disabled (Gemini `thinking_budget=0`); "dynamic" = model sets its own reasoning depth (`thinking_budget=−1`); "on" = reasoning always on with no depth control; "minimal"/"high" = named reasoning-effort tier. Two Gemini IDs are aliases; at run time (July 2026) `gemini-pro-latest` resolved to `gemini-3.1-pro-preview`, and `gemini-3.5-flash` was pinned explicitly rather than `gemini-flash-latest` (which then resolved to `gemini-3.6-flash`). [a]Claude rejects `temperature`/`top_p` and exposes no seed; extended thinking was not enabled. [b]Inkling accepts a seed but the API echoed `null`, so determinism is best-effort. [c]Muse Spark exposes no temperature or seed control, and its reasoning depth is non-deterministic run-to-run. [d]`gemini-pro-latest` cannot disable thinking and always reasons dynamically.

Modality	α	Pearson r	p	Spearman ρ
Text	0.91	0.29	.004	0.23
Audio	0.88	0.03	.78	0.00
Video	0.92	−0.12	.24	−0.19

Table A3: SP subscale reliability and its correlation with per-rater accuracy ($N = 94$ text, 94 audio, 92 video).

($r = 0.45$, $p < .001$), with weak raw associations in audio ($r = 0.19$, $p = .06$) and text ($r = -0.04$, n.s.). Confidence-weighted single-model estimates match the binary ones almost exactly, so this is not an artifact of one prediction per dyad. Individual models thus express the human-aligned difficulty signal only noisily, and pooling recovers it.

Estimator	Text	Audio	Video
Pooled, raw	0.20	0.56	0.34
Pooled, within-class	0.61	0.58	0.44
Best model, raw	−0.04	0.19	0.45
Best model, within-class	0.24	0.26	0.47

Table A4: Human–model per-dyad difficulty correlation (r) under four estimators, by modality ($N = 96$ dyads). "Pooled" averages model accuracy over all models; "best model" is the single most accurate model per modality (text `gpt_text_mini`, audio `gemini_audio_pro`, video `gemini_video`). "Within-class" partials out the true familiar/stranger label.

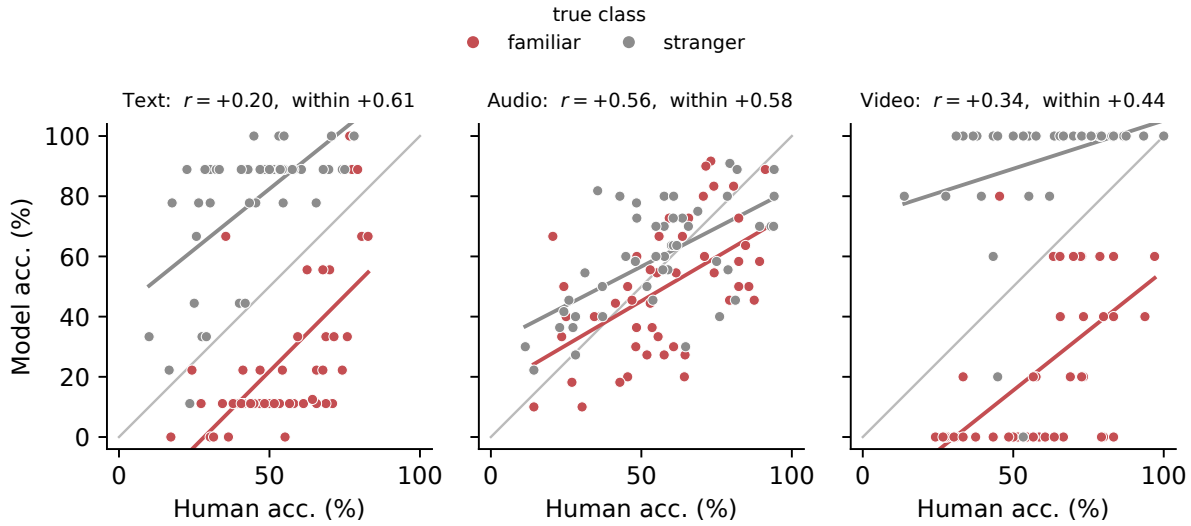


Figure A1: Humans and models tend to find the same dyads hard (Section 5.5). Each point is one of the 96 EO dyads, plotting human per-dyad accuracy (over all raters) against model per-dyad accuracy (pooled over all models), colored by true class; the diagonal marks equal difficulty. Per-dyad difficulty correlates positively in audio ($r = 0.56$, $\approx 31\%$ shared variance) and video ($r = 0.34$, $\approx 12\%$; both $p < .01$); text ($r = 0.20$) is only marginally significant ($p = .045$).

Dyad	Truth	Txt.	Aud.	Vid.	Mean
Hardest					
`sample_080`	str.	29.6	32.0	22.2	27.9
`sample_055`	str.	20.5	32.1	33.3	28.6
`sample_072`	fam.	35.4	19.9	38.0	31.1
`sample_053`	str.	31.1	27.0	40.7	32.9
`sample_038`	fam.	25.8	28.5	47.9	34.1
`sample_057`	str.	34.1	30.3	37.9	34.1
`sample_083`	str.	31.7	34.7	36.1	34.2
`sample_065`	fam.	34.9	33.4	35.0	34.4
Easiest					
`sample_045`	fam.	59.6	83.8	79.3	74.2
`sample_062`	str.	69.2	75.2	79.4	74.6
`sample_051`	str.	53.3	88.4	82.0	74.6
`sample_077`	fam.	57.4	79.1	89.9	75.5
`sample_009`	fam.	70.7	77.5	87.4	78.5
`sample_030`	str.	63.3	83.7	91.1	79.4
`sample_073`	fam.	71.4	86.1	81.9	79.8
`sample_016`	str.	68.5	86.8	86.8	80.7

Table A5: Hardest and easiest eight dyads by mean estimated P(correct human response) across modalities (%). "str." = stranger, "fam." = familiar.

MentaLED: A Multimodal Conversational Agent for Mental Health Support Using Linguistic and Paralinguistic Speech Cues

Changhyun Kim[1,*,†], **Dahye Kim**[2,*], **Eunseon Seong**[2,*], **Harim Lee**[2,*], **Dong-Kyu Chae**[2] and **Sangyool Lee**[1]
[1]Network AX CoE, SK Telecom, Seoul, South Korea
[2]Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[*]Equal contribution.
[†]Corresponding author: changhyk@sktelecom.com

Abstract

Mental health, defined by psychological, emotional, and social well-being, has emerged as a significant social issue in modern society. To assess an individual's mental state and provide personalized support, clinician–patient interviews are commonly conducted. However, given the constraints on access to such interviews, conversational agents have the potential to provide more accessible mental health support. In this paper, we present MentaLED, a multimodal conversational agent that assesses an individual's emotional and depressive states using an audio–text model that jointly models linguistic and paralinguistic cues from speech, and provides personalized mental health support through LLM-based response generation. Quantitative and qualitative evaluations indicate improvements in multimodal state assessment and response quality over the considered baselines.

1 Introduction

Mental health is a fundamental aspect of overall well-being, encompassing psychological, emotional, and social well-being (Galderisi et al., 2015; Seong et al., 2026). The World Health Organization (WHO) emphasizes its importance with the proposition of "no health without mental health." In clinician–patient interviews, speech provides rich information not only about what a person says, but also about how it is expressed through paralinguistic cues such as tone, pitch, and intonation. However, access to such interviews is often constrained by limited clinical resources, financial costs, and scheduling constraints, motivating automated approaches for mental health support. Recent LLM-based methods have shown promise in emotional support (Zheng et al., 2023), but text-only systems cannot directly exploit paralinguistic information from speech.

Prior studies have explored depression detection from speech using hand-crafted acoustic and prosodic features and, more recently, deep learning, while broader studies have identified both acoustic and linguistic speech characteristics as promising biomarkers for depression (Cummins et al., 2011; Koops et al., 2023; Wu et al., 2023; Maran et al., 2025). More recently, emotion2vec (Ma et al., 2024) demonstrated that self-supervised pre-training can learn generalizable speech representations for emotion recognition across diverse datasets, languages, and emotion-related tasks. However, these lines of work largely treat emotion or depression recognition as standalone representation or prediction tasks, without connecting the inferred mental state to downstream personalized conversational support. In contrast, MentaLED connects joint linguistic–paralinguistic state inference directly to LLM-based response generation, forming a unified pipeline from speech-based mental-state assessment to personalized conversational interaction.

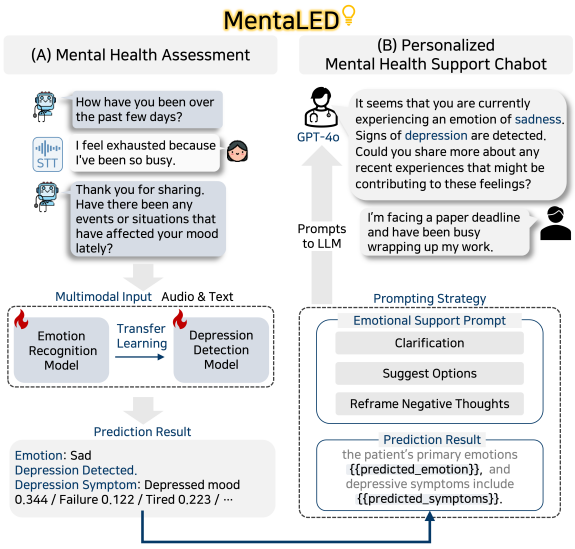


Figure 1: Overall pipeline of MentaLED.

To this end, we propose **MentaLED**, a multimodal conversational agent for personalized mental health support. MentaLED operates in two steps:

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 144–149
October 29, 2026 ©2026 Association for Computational Linguistics

No.	Question
1	How have you been over the past few days? Have you felt any particular feelings recently?
2	Thank you for sharing, it sounds like this has been meaningful for you. Have there been any specific events or situations that have affected your mood lately?
3	When you feel this way, what kinds of thoughts usually come to mind?
4	Have you been experiencing any changes in appetite, sleep, or energy levels recently?
5	Would you say these feelings have been consistent, or have they occurred occasionally?

Table 1: Question set for mental health assessment.

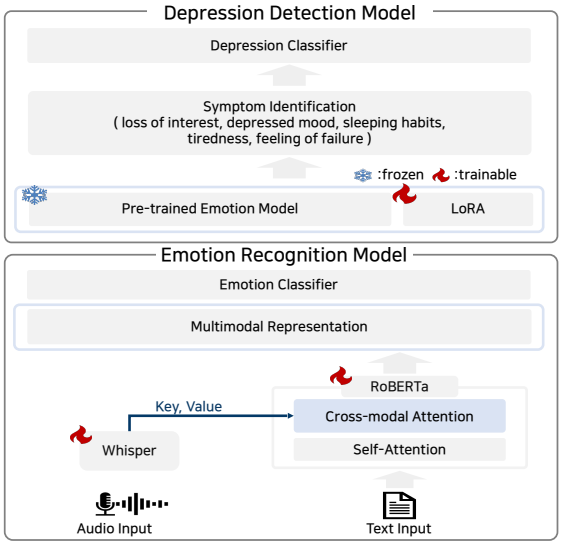


Figure 2: Model Architecture.

1) Emotional and Depressive State Assessment and 2) Personalized Response Generation based on the assessed state, as illustrated in Figure 1.

For state assessment, we first acquire the user's spoken responses to our question set through a microphone. Speech provides two complementary sources of information: linguistic cues that convey users' experiences, thoughts, and concerns, and paralinguistic cues, such as tone, pitch, and intonation, that reflect how those experiences are expressed. To jointly model these cues, we introduce a novel multimodal (audio and text) model that integrates audio and text representations. Specifically, we employ cross-modal attention to derive an audio-referred text representation that encodes both linguistic and paralinguistic cues in speech, which are then used to predict the user's emotional and depressive states.

Based on the inferred emotional and depressive states, we leverage an LLM to generate personalized mental health support through a conversational agent. Through extensive evaluation, we demonstrate the effectiveness of MentaLED in providing effective and appropriate mental health support.

Prompt

You are a mental health counselor specializing in providing emotional support and guidance to patients. Your task is to assess the emotions and mental state of the patient and offer counseling accordingly, following the structured emotional support strategies listed below.
Clarification: Always seek to understand the patient's emotions or experiences more clearly. Ask additional questions when needed to ensure a comprehensive understanding of their current emotional state.
Suggest Options: When the conversation progresses, and you have a deeper understanding of the patient's situation, provide practical suggestions or alternative ways to address their current challenges.
Reframe Negative Thoughts: If the patient expresses particularly negative or harmful thoughts, help them reframe those thoughts into more constructive or realistic perspectives.

Here's the information you are provided with: the patient's primary emotion is **{predicted_emotion}**, depressive symptoms include **{depression_symptoms_str}**, and depression status is **{depression_status}**.

The counseling process will be divided into three phases:
1. **Initial Phase**: Focus on asking questions to clarify the patient's current emotional state and offer empathy. Use the **Clarification** strategy to ensure you understand their feelings clearly.
2. **Middle Phase**: Begin to ask more specific questions about their symptoms to explore the root causes. Continue to validate their feelings and, if appropriate, start suggesting some **Suggest Options** for them to consider.
3. **Final Phase**: Based on the patient's depressive symptoms and your analysis, provide practical solutions using the **Suggest Options** strategy. If the patient expresses negative self-perception, employ the **Reframe Negative Thoughts** strategy to help them adopt a more positive outlook.
Based on the patient's depressive symptoms and your analysis, provide practical solutions using the strategy.

Table 2: Prompt for mental health support response generation.

2 MentaLED

MentaLED consists of two components: (A) Multimodal Emotion Recognition and Depression Detection, and (B) Response Generation for Mental Health Support.

Multimodal Emotion Recognition and Depression Detection. For emotion recognition, we use ***IEMOCAP***, a multimodal conversational dataset with six emotion labels: happy, sad, angry, excited, frustrated, and neutral. For depression detection, we use ***DAIC-WOZ***, which contains 189 clinical interviews annotated for depression and eight depressive symptoms (Gratch et al., 2014). Following prior work (Agarwal et al., 2024), we select five symptoms that can be inferred from conversational responses: loss of interest, depressed mood, sleeping problems, tiredness, and feelings of failure.

Next, we propose a novel multimodal model in-

tegrating audio and text modalities to capture the rich information embedded in the user's speech, as illustrated in Figure 2. While the audio modality conveys paralinguistic cues, including non-verbal vocal expressions (e.g., tone, pitch, intonation), the text modality captures contextual information. To integrate these different modalities effectively, we use the text as the primary modality to extract semantic features and leverage the audio modality as complementary information to provide additional cues for emotion. Specifically, we adopt well-known models based on attention mechanisms, RoBERTa (Liu, 2019) for the text modality and Whisper (Radford et al., 2023) for the audio modality. To obtain a multimodal representation, we introduce a cross-modal attention layer after the self-attention layer (Li et al., 2021). In each cross-modal attention layer, the text stream serves as the queries, while the final audio representations from Whisper serve as the keys and values. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is applied to achieve parameter efficiency.

The proposed model is trained in two stages, where we first train the emotion recognition model based on supervised learning. Then, based on the finding that depression is strongly associated with emotional states (Compare et al., 2014), we apply transfer learning for depression detection by utilizing the emotion recognition model as a pre-trained model. Since various depressive symptoms are involved in depression, the depression detection model comprises symptom identification followed by overall detection prediction.

Response Generation for Mental Health Support. We leverage GPT-4o to generate responses conditioned on the assessed mental health state, with two prompting strategies as illustrated in Table 2. First, we adopt the user's emotional state and depressive symptoms as prompts, enabling the LLM to generate responses tailored to the individual's mental health condition.

Additionally, we incorporate emotional support strategies inspired by ExTES (Zheng et al., 2023), which introduces prompting techniques for empathic conversation. Specifically, we adopt three approaches to guide the response generation process in a structured sequence and provide personalized psychological support: **(1) Clarification** to better understand the user's emotions and experiences, **(2) Suggesting options** to provide practical recommendations or alternative solutions to address the user's challenges, and **(3) Reframing negative thoughts** to help users reinterpret harmful or distressing thoughts constructively. We select these strategies to translate the assessed emotional and depressive states into context-sensitive support. Clarification explores events that triggered sadness, suggesting options introduces coping techniques such as relaxation exercises, and reframing negative thoughts redirects self-critical judgments toward what the user has learned from the experience. This component is integrated into the conversational agent.

3 Experiments

To evaluate performance, we compare MentaLED against two baselines: 1) ***None***: a general LLM (GPT-4o) without any prompts or user state consideration, and 2) ***ExTES***: an emotional support conversational agent using three diverse prompting strategies adopted from ExTES (Zheng et al., 2023).

3.1 Emotion Recognition and Depression Detection

Table 3 presents the results of the emotion recognition task on IEMOCAP and the depression detection task on DAIC-WOZ. We use the official train/test splits provided by the respective benchmarks. We report the weighted F1 score and accuracy for each model configuration. 'Audio-only' and 'Text-only' refer to unimodal variants using only the audio representation from Whisper (Radford et al., 2023) and the text representation from RoBERTa (Liu, 2019), respectively. 'Multimodal' represents the proposed model, which integrates both modalities through the cross-modal attention mechanism as described in Section 2. Across all three settings, the same emotion classifier architecture is used, such that the configurations differ only in how the representations are obtained. Our model achieves the highest performance, confirming its effectiveness in extracting multimodal features.

Task	Dataset		Modality		
			Audio-only	Text-only	**Multimodal**
Emotion	IEMOCAP	WF1	53.79	56.39	**61.71**
		ACC	54.02	57.30	**62.05**
Depression	DAIC-WOZ	WF1	57.51	56.61	**59.84**
		ACC	60.60	57.58	**63.63**

Table 3: Performance of our emotion recognition and depression detection models under different settings.

3.2 Mental Health Support

To evaluate the appropriateness of the generated mental health support across different emotional and depressive states, user responses under various conditions are required. For this purpose, we leverage an LLM to generate user responses reflecting one of 6 emotional states and one of 5 depressive states (30 distinct states) and evaluate the generated response to each simulated user response.

(a) Results of emotional support assessment

	Task 1		Task 2		Task 3		Task 4	
	PASS	WIN	PASS	WIN	PASS	WIN	PASS	WIN
None	77.33	**86.67**	73.33	83.33	86.67	80.00	80.00	90.00
ExTES	83.33	80.00	83.33	**96.67**	**90.00**	90.00	90.00	86.67
MentaLED	**96.67**	83.33	**86.67**	**96.67**	**90.00**	**96.67**	**100.00**	**96.67**

(b) Results of mental health assessment

	Question 1	Question 2	Question 3	Question 4
None	8.277	6.744	5.422	8.078
ExTES	8.289	6.655	5.189	7.877
MentaLED	**8.544**	**6.989**	**6.533**	**8.377**

Table 4: Evaluation of mental health support responses.

We conduct two assessments: Emotional Support Assessment to verify the conversational agent's comprehensive empathy capability (Chen et al., 2024) and Mental Health Support Assessment to evaluate the efficacy in delivering meaningful mental health support, especially in terms of safety and reliability. For both assessments, we utilize 3 LLM judges as evaluators - GPT-4o, Claude-3.5, and Mistral-Large. Note that the reported results for each assessment represent the average across 30 distinct states and all LLM judges.

Emotional Support Assessment. To assess the comprehensive empathy capability of the conversational agent, we utilize an evaluation benchmark, EmotionQueen (Chen et al., 2024). EmotionQueen proposes 4 tasks to capture diverse aspects of emotional understanding: ***1. Key Event Recognition*** (identification of the most significant emotional event), ***2. Mixed Event Recognition*** (recognition of multiple emotional events reflecting emotional complexity), ***3. Implicit Emotion Recognition*** (detection of underlying deep emotions), ***4. Intention Recognition*** (understanding of potential purposes or intents and providing appropriate support) and two metrics (PASS, WIN rate) to evaluate the performance in recognizing and responding to emotion-related statements. The PASS rate measures the accuracy of emotional recognition, while the WIN rate evaluates the ability to generate empathetic responses. The LLM judges assign 0 or 1 to each metric, with a score of 1 indicating better performance.

Table 4(a) shows the PASS and WIN rates, where MentaLED achieves the highest PASS scores across all tasks, demonstrating accurate emotion recognition. In terms of WIN scores, MentaLED outperforms the baselines overall, particularly with a large margin in Tasks 3 and 4, which require deep emotional understanding and contextually appropriate generation. The overall result highlights the superiority of MentaLED in understanding emotional states and generating empathetic, contextually appropriate responses.

Mental Health Support Assessment. To validate the conversational agent's reliability in delivering meaningful mental health support, we employ the evaluation method proposed in prior work (Park et al., 2024). This work proposes four guideline questions to evaluate the diverse safety aspects of a mental health conversational agent: ***1. Adherence to Practice Guidelines*** (adherence to clinical safety standards and established practice guidelines), ***2. Identification and Management of Health Risks*** (recognition and appropriate response to potentially dangerous health situations), ***3. Resource Provision Assessment*** (delivery of relevant external support resources), ***4. User Empowerment for Health Management*** (encouragement of proactive self-care while emphasizing professional oversight).

Table 4(b) presents the scores for each guideline question on a 10-point scale, where MentaLED achieves the highest scores across all criteria. Our preliminary results suggest that MentaLED can provide contextually appropriate and supportive responses; however, these findings should not be interpreted as evidence of clinical efficacy or safety.

3.3 Qualitative Results

We provide a qualitative comparison to demonstrate how MentaLED generates responses tailored to the user's mental health state. Figure 3 presents responses of MentaLED alongside those of None (GPT-4o) and ExTES in the same scenario. While None and ExTES provide suggestions based solely on the user's input, without consideration of the underlying psychological state, MentaLED leverages both the user's input and the predicted emotional and depressive symptoms to provide personalized mental health support. For example, when a user mentions "I became lethargic because I was busy and kept working," MentaLED interprets this within the context of the user's assessed emotional

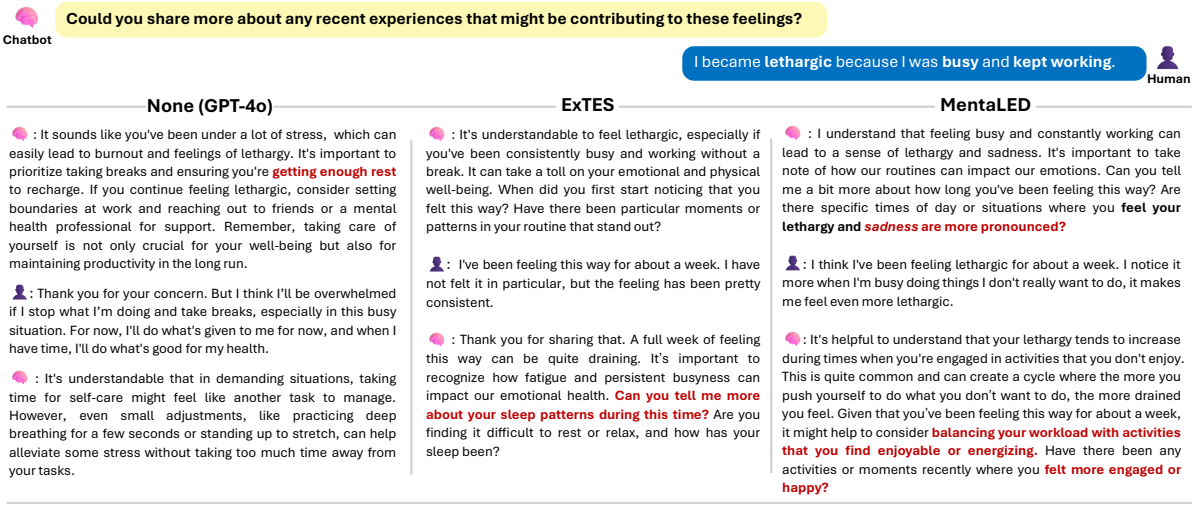


Figure 3: Qualitative comparison of model responses.

state (sadness) and recommends energizing or enjoyable activities to address the underlying situation. This distinguishes MentaLED from the baselines, where None and ExTES suggest general solutions (e.g., improving sleep habits, taking a rest) that have limited ability to address the user's psychological needs.

3.4 Human Evaluation

To assess the practical usability of the prototype, we conduct a small-scale human evaluation with 10 participants. Each participant responds to the question set in Table 1 through a microphone and interacts with the conversational system, which generates support based on the state predicted by our model. Participants then compare responses from the conversational agent models in a pairwise manner according to four criteria: **(Q1) Appropriate guidance**, **(Q2) Situation-specific counseling**, **(Q3) Detailedness**, and **(Q4) User preference**.

Table 5 reports the proportion of pairwise comparisons in which each model is preferred. MentaLED achieves the highest preference across all criteria, suggesting that conditioning responses on emotional and depressive states can improve perceived response quality. This evaluation measures user-perceived response quality rather than clinical effectiveness or diagnostic validity.

	Question 1	Question 2	Question 3	Question 4
None	0.00	0.00	0.00	0.00
ExTES	0.44	0.33	0.44	0.44
MentaLED	**0.56**	**0.67**	**0.56**	**0.56**

Table 5: Human evaluation results.

4 Conclusion

We present MentaLED, a multimodal conversational agent that integrates emotion recognition and depression detection with LLM-based response generation. By jointly modeling linguistic and paralinguistic speech cues, MentaLED infers emotional and depressive states and uses them to generate state-aware support responses. Experimental and human evaluation results indicate improvements over the considered baselines in multimodal state assessment and perceived response quality. These findings represent preliminary evidence of system-level effectiveness rather than clinical validation.

Limitations

Although MentaLED integrates emotion recognition and depression detection with personalized response generation, several limitations remain. First, the system has not been clinically validated or reviewed by licensed mental health professionals. Safety issues including crisis handling, privacy, and deployment constraints also require further investigation. Second, LLM-based judges may introduce model-specific biases, and differences between the LLM and human evaluation criteria limit direct comparison between them. Third, our human evaluation is small-scale and assesses prototype interaction and pairwise response preference rather than clinical outcomes. Further evaluation using shared criteria, larger participant groups, and mental health professionals is therefore needed. Our preliminary results should not be interpreted as evidence of clinical efficacy or safety.

Acknowledgement

We would like to acknowledge that this work mainly benefited from **mentorship provided through the SK Telecom AI Fellowship Program**. This work was partly supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (**RS-2024-00413747** and **RS-2024-00413585**) and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(**RS-2020-II201373**, Artificial Intelligence Graduate School Program (Hanyang University)).

References

Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. 2024. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983.

Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*.

Angelo Compare, Cristina Zarbo, Edo Shonin, William Van Gordon, and Chiara Marconi. 2014. Emotional regulation and depression: A potential mediator between heart and mind. *Cardiovascular psychiatry and neurology*, 2014(1):324374.

Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Interspeech 2011*, pages 2997–3000.

Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. 2015. Toward a new definition of mental health. *World psychiatry*, 14(2):231.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sanne Koops, Sanne G. Brederoo, Janna N. de Boer, Femke G. Nadema, Alban E. Voppel, and Iris E. Sommer. 2023. Speech as a biomarker for depression. *CNS & Neurological Disorders – Drug Targets*, 22(2):152–160.

Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Patricia Laura Maran, María Dolores Braquehais, Alexandra Vlaic, María Teresa Alonzo-Castillo, Júlia Vendrell-Serres, Josep Antoni Ramos-Quiroga, and Amanda Rodríguez-Urrutia. 2025. Performance of automatic speech analysis in detecting depression: Systematic review and meta-analysis. *JMIR Mental Health*, 12:e67802.

Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn Bounds, Angela Jun, Jaesu Han, Robert McCarron, Jessica Borelli, Jia Li, Mona Mahmoudi, et al. 2024. Building trust in mental health chatbots: safety metrics and llm-based evaluation tools. *arXiv preprint arXiv:2408.04650*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Eunseon Seong, Harim Lee, Dahye Kim, Changhyun Kim, and Dong-Kyu Chae. 2026. Emotion-wheel-guided audio-referred text representation for multimodal emotion recognition in conversation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40392–40403.

Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. 2023. Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, 8(3):701–711.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Annotating and Recognizing the Palm Up Open Hand Gesture in Dutch

Fenna Berger, Alan Cienki, Lucia Donatelli
Vrije Universiteit Amsterdam / Amsterdam, Netherlands
f.berger@student.vu.nl, l.e.donatelli@vu.nl

Abstract

Recurrent gestures are conventionalized gestures with a stable relationship between form and meaning. One such gesture is the palm up open hand (PUOH) gesture, typically produced by extending the fingers and rotating the lower arm so that the palm faces upward. The gesture is attested across many languages and cultures and is often used to present a point, yet little is known about the forms and functions it takes in Dutch. Automatic gesture recognition could reduce the manual labor that quantitative gesture studies require, but recognizing this gesture is difficult, partly because of the large amount of variation present in its form. This paper compares two gesture recognition approaches on a newly collected corpus of naturalistic Dutch dyadic conversation: a static model that takes single images as input and a dynamic model that takes video clips as input. Both models are trained and tested with varying amounts of gesture-form variation included in the data. The static model is found to be less affected by this variation and to yield overall better results than the dynamic model. Our results illustrate some of the central challenges that gesture recognition systems face in recognizing the PUOH gesture and offer insights that may improve performance in future studies of multimodal, face-to-face behavior.

1 Introduction

Gestures are an important part of face-to-face communication. They carry pragmatic and semantic information that adds a further layer of meaning to spoken language, and are attested across cultures (Kendon, 2004; Kita, 2009), ages (Feyereisen and De Lannoy, 1991), and communicative contexts (McNeill, 1992; Chu et al., 2014), including in children at the one-word stage (Iverson and Goldin-Meadow, 2005) and in congenitally blind speakers talking to blind listeners (Iverson and Goldin-Meadow, 1998; Chu et al., 2014).

Recurrent gestures are one gesture type that has only recently attracted sustained attention in gesture studies: a gesture is recurrent if it "is used repeatedly in different contexts and its formational and semantic core remains stable across different contexts and speakers" (Müller, 2010, p. 41). Qualitative work on recurrent gestures exists (Ladewig, 2014; Bressem and Müller, 2014), but quantitative work is rare, as it typically requires large amounts of costly, time-consuming manual annotation.

A possible solution is to use automatic gesture recognition models to support annotation, for instance by flagging candidate gestures for human annotators to label. However, such models are difficult to train, in part because of the high degree of variation present within and across individuals producing the same gesture; a recognition model must generalize across this variation to classify gestures correctly.

This paper investigates the broad question of how gesture recognition models can recognize recurrent gestures in Dutch, focusing on the palm up open hand (PUOH) gesture, in which the speaker rotates the lower arm to present the palm to an interlocutor, often to make a point. A preliminary analysis of our newly collected corpus of Dutch dyadic conversation found this gesture to be highly frequent, consistent with prior qualitative work showing it is used across spoken and signed languages (Müller, 2004; Cooperrider et al., 2018), but no existing work was found on the forms this gesture takes specifically in Dutch. This motivates two more specific research questions:

> **RQ1**: What forms does the palm up open hand recurrent gesture take in Dutch?
>
> **RQ2**: How can gesture recognition models recognize these gestures, in particular with respect to (a) individual variation and (b) limited training data?

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 150–159
October 29, 2026 ©2026 Association for Computational Linguistics

This work is exploratory: although the PUOH gesture is among the most extensively studied recurrent gestures qualitatively, no prior study was found that attempts to detect it automatically with machine learning. Given this, and given the small size of the available data set, the goal is not to build the best-performing model, but to compare two model types with a focus on understanding how they work, which factors affect their performance, and which challenges they face.

One such factor that may impact performance is variation in gesture form. The different forms of the PUOH gesture lie on a continuum, ranging from a small finger lift to an extension of the entire arm (Cienki, 2021). A model must generalize across these variants to recognize the gesture reliably. A second factor is the limited availability of training data, which typically induces class imbalance, since speakers spend far more time not producing a PUOH gesture than producing one. Restricting the positive class to fewer form variants reduces this imbalance but further limits the amount of positive training data, creating a trade-off that this work investigates directly.

The contributions of this paper are threefold: (1) a small annotated corpus of naturalistic Dutch dyadic conversation, part of which will be made available to other researchers upon request, in line with participant consent; (2) an empirical characterization of the forms the PUOH gesture takes in Dutch (RQ1); and (3) two gesture recognition models, trained and evaluated across conditions that vary in form-variation and data-imbalance, together with an analysis of the trade-offs between them (RQ2).

2 Related Work

Recurrent gestures. Interest in recurrent gestures is a relatively recent development within linguistics, although related phenomena were long studied under different names in rhetoric and acting (Loeb, 1912; Mosher, 1916; Ott, 1902). Early linguistic treatments used terms such as *pragmatic gestures* (Kendon, 1995; Streeck, 2005, 2009) or *interactive gestures* (Bavelas et al., 1992, 1995); Ladewig (2014) argued these terms fail to capture the language-like properties of such gestures, and the term *recurrent gesture* was later coined within the "Towards a Grammar of Gesture" project (Müller et al., 2013). Recurrent gestures resemble emblematic gestures in showing a stable form–meaning relationship, but unlike emblems they always act on or accompany a verbal utterance rather than replacing it (Bressem and Müller, 2014). Recurrent gestures can serve either a referential function (depicting objects, actions, or events, typically co-expressed with speech) or a meta-communicative function as a speech performative, acting on the discourse structure or expressing an attitude (Ladewig, 2014; Teßendorf, 2014).

The palm up open hand (PUOH) gesture. The PUOH gesture combines a rotation of the lower arm, so the palm faces upward, with an extension of the fingers, and occurs frequently across spoken and signed languages (Cooperrider et al., 2018). Müller (2004) described it as metaphoric, rooted in the practical actions of giving and receiving, with giving-related senses (e.g., presenting an object for joint inspection) and receiving-related senses (e.g., expressing not knowing). Kendon (2004) termed it the "open hand supine" and distinguished palm presentation, palm addressed, and palm lateral variants, associated respectively with explaining or summarizing, acknowledging or offering, and expressing unwillingness, obviousness, or a rhetorical stance. Quantitative studies remain rare: Chu et al. (2014) found that palm-revealing and conduit gestures (subtypes of PUOH) made up about 24% of all gestures produced while participants defined phrases and solved social dilemmas, and that their frequency correlated with participants' empathy scores. Jehoul et al. (2017), analyzing roughly eight hours of Flemish Dutch conversation, found the palm lateral variant to be the most frequent marker of obviousness, supporting Kendon's account. No existing work was found on which PUOH forms occur in Dutch specifically, which the present corpus addresses.

Gesture recognition. Automatic recognition is one way to reduce the annotation cost that limits quantitative gesture research (Xu, 2017; Zou et al., 2018; Tateno et al., 2019; Ojeda-Castelo et al., 2022). Within hand-gesture recognition, vision-based approaches (Sinha et al., 2019) are more common than sensor-based (glove) approaches (Kim et al., 2019), since gloves are more invasive and may alter how participants gesture. A further distinction is between *static* recognition, which classifies poses (commonly using SVM, k-NN, or clustering methods; Cheok et al., 2019; Tharwat et al., 2015; Kurdyumov et al., 2011), and *dynamic* recognition, which classifies movements over time

(commonly using dynamic time warping or Hidden Markov Models; Al-Shamayleh et al., 2018; Darrell and Pentland, 1993; Chen et al., 2003). Recognized challenges in this field include temporal variance (the same gesture produced at different speeds), spatial complexity (the same gesture produced at different sizes, locations, or orientations), sensitivity to camera angle, the cost of annotated training data, and the difficulty of segmenting gestures that occur in continuous, sometimes cyclic sequences (Nam and Wohn, 1996).

Variation as signal, not noise. Both linguistics and NLP have tended to treat data variation as noise to be eliminated, despite variation being a core feature of language use through which speakers express identity and social affiliation (Curry et al., 2024). Schmidt (2010) distinguishes micro- (individual), meso- (group), and macro- (standard-language) levels of linguistic "variation space," and observes that NLP typically models only the macro-level, treating other variation as disruptive (Plank, 2016). Lutgen et al. (2026) argue that data selection should be guided by which varieties should be represented rather than by data volume alone, illustrated by Ramponi (2024)'s speaker-centric critique of NLP for non-standard Italian. In the specific case of PUOH gestures, Cienki (2021) argues that existing gesture-description systems based on external features (hand shape, palm orientation, direction and manner of motion; Stokoe Jr, 1960/2005; McNeill, 1992; Bressem et al., 2013) fail to capture the continuum linking different PUOH forms, and proposes instead the internal, joint-based kinesiological system of Boutet et al. (2018), which describes gestures in terms of flexion/extension, abduction/adduction, rotation, and supination/pronation at each joint. This system is adopted for the annotation scheme used in the present study (Section 3).

3 Data

3.1 Collection

Data collection followed the methodology of Cienki et al. (Forthcoming) (briefly described in Cienki et al., 2025; Prokofyeva et al., 2025) and was approved by the host institution's Research Ethics Review Committee. Recordings took place over three days in classrooms with a neutral background, with two participants seated at a 90 angle in front of a table holding a laptop. Three GoPro cameras (two HERO12 Black, one HERO9 Black) were used (one overview camera and one individual camera per participant) together with a lapel microphone per participant (Fig. 1).



Figure 1: Schematic representation of the setup used for data collection as described in Sec. 3.

Ten pairs of native Dutch-speaking students who already knew each other (N=20, 13 female) took part, recruited on a university campus and compensated with a gift card. Each pair first watched a short video about artificial intelligence (AI)(with a Dutch voiceover) as a shared conversational prompt, then discussed a series of questions for up to 20 minutes. Gestures were not mentioned to participants at any point, to avoid drawing attention to them. Minor data-collection issues occurred, including one camera losing roughly half of one session's recording due to an SD-card fault, a camera tilting out of position on the first day, and a brief interruption to charge equipment; two conversations were briefly interrupted by the experimenter for unrelated reasons (a participant consulting a phone, and a participant becoming distracted by suspicion about the study's purpose).

3.2 Annotation

Recordings were synchronized in Adobe Premiere Pro and annotated in ELAN (Max Planck Institute for Psycholinguistics, The Language Archive, 2026), marking the full gesture (stroke, hold, retraction) where possible. Annotations were divided into a positive class (PUOH gestures) and a negative class (all other (non-)gestures), each with four subclasses. Following the joint-based description in Cienki et al. (2025), positive-class instances were labeled by which parts of the arm and hand were involved: thumb lift (TL; thumb only), hand turnout (HT; wrist rotation), open hand diagonal (OHD; elbow rotation/extension), and palm up open hand (PUOH; shoulder rotation/extension), treated as increasingly larger points on Cienki's continuum. Negative-class instances were labeled as other gesture (OG), self-adapter (SA; e.g., fidgeting, adjusting hair), no gesture (NG), or unsure (US; gestures resembling PUOH in form but



Figure 2: Example image of a participant making a positive-class gesture (affirmative PUOH; Cienki et al. (2025)) as used in the static model data.



Figure 3: Example image of a participant making a negative-class gesture (non-PUOH) as used in the static model data.

judged, by native-speaker intuition, to serve a different function, e.g. iconic). For the static model, one image was manually captured at the hold of each gesture (or, where no clear hold existed, at the point where the hands were most in focus); for the dynamic model, video clips were extracted automatically from the annotation boundaries, with individual per-participant videos trimmed from the overview and side cameras and mirrored for right-seated participants to control for handedness effects related to seating position. Figure 2 shows an example of a screenshot of a positive-class instance used for the static model. Figure 3 shows the same for a negative-class instance.

The full corpus comprises 3h04m40s of recording, 1,812 annotated instances (567 positive, 1,245 negative; Tables 1 and 2), with positive clips averaging 1.58s and negative clips 1.90s in duration. Within the positive class, open hand diagonal was the most frequent subclass (298 instances) and thumb lift the least frequent (12); within the negative class, other gesture was most frequent (580) and unsure least frequent (139). The number of gestures produced varied considerably across participants (22–197 total gestures, excluding the session with the camera fault; average 98.17), as did the proportion that were positive-class gestures (0.05–0.50, average 0.29), consistent with known individual variation in gesture use (Chu et al., 2014); no instances were excluded from the data set on this basis, since such variation reflects genuine differences among language users.

Class	Subclass	Total
Positive	Thumb lift	12
	Hand turnout	114
	Open hand diagonal	298
	Palm up open hand	143
	Total	567
Negative	Other gesture	580
	Self-adapter	247
	No gesture	279
	Unsure	139
	Total	1245
Overall	Total	1812

Table 1: Total number of occurrences of classes and subclasses of palm up open hand (PUOH) gestures. Positive indicates affirmative PUOH gesture; negative indicates another gesture type.

	Min	Max	Avg	Total
Recording length	10:20	20:19	18:28	3:04:40
Positive instances	4	135	57	567
Negative instances	41	190	125	1245

Table 2: Corpus statistics across ten recording sessions (min/max/avg per session, and totals). Instances refer to PUOH gestures.

4 Models

Two gesture recognition approaches were compared: a *static* model, which classifies single images, and a *dynamic* model, which classifies video clips. Because this project was conducted within the scope of a master's thesis, the model selection was heavily influenced by the limited time and resources available. While this means that the results, do not reflect the abilities of state-of-the-art gesture detection models, they do shine light on some differences between gesture detection models, as well as challenges that need to be tackled in the future. All code will be made publicly available upon acceptance.

4.1 Static Model

The static model adapts Tokhirov (2023), who trained several classifiers on MediaPipe hand landmarks (21 joint coordinates per hand; Fig. 4) to

distinguish four webcam-recorded hand shapes, one of which closely resembles the PUOH hand shape (though facing the camera rather than upward). Of the three classifiers Tokhirov compared (a dense neural network, XGBoost, and random forest), the random forest performed best while also being the fastest, and is adopted here, retrained as a binary classifier on the present, more naturalistic data (which, unlike Tokhirov's, contains variation in gesture form, viewing angle, lighting, and 20 distinct speakers).

4.2 Dynamic Model

The dynamic model adapts the convolutional architecture of Pouw (2024), itself based on Yung (2022), which uses MediaPipe Holistic to extract 29 features per 25-frame window (16 hand-related, plus rotation, facial, and body-distance features, together with their time-derivatives and standard deviations). The network has three convolutional layers (32, 64, and 128 filters) with pooling between layers, and outputs a probability of motion for each window together with probabilities for each gesture type; following Pouw (2024), a motion threshold of 0.5 and a gesture threshold of 0.9 are used to assign labels. Two changes were made to the original design: the model was trained for 5 rather than 500 epochs, owing to computational constraints (a limitation discussed in Section 8), and, since one label per clip (rather than per window) was required, the highest per-clip motion and PUOH probabilities were used to assign the clip-level label.

5 Experimental Setup

Both models were trained and tested with a 70/30 split, balanced across the eight subclasses. Conditions varied the subclasses included in the positive class (six combinations for the static model: PUOH-only, OHD-only, HT+TL-only, PUOH+OHD, OHD+HT+TL, and All-Variants, merging TL with HT owing to its low frequency) and in the negative class (with or without the Unsure subclass), to test how increasing variation and class imbalance affect performance. Owing to its longer training time, the dynamic model was evaluated on a subset of four conditions (OHD-only and All-Variants, each with and without Unsure). A follow-up experiment repeated the static-model conditions with a fixed sample size (250 images per class) to disentangle the effect of variation from

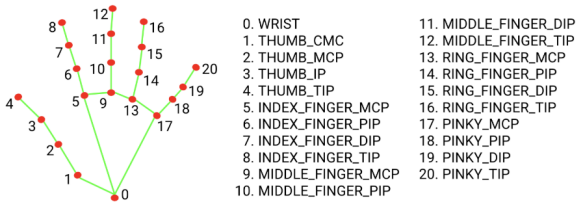


Figure 4: Overview of the 21 landmarks that are tracked by MediaPipe handtracking (source).

that of sample size.

Models were evaluated using accuracy, precision, recall, and macro F1 (the unweighted mean of per-class F1, giving equal weight to each class regardless of size), since accuracy alone can be misleadingly high under class imbalance.

6 Results

6.1 Static Model

Table 3 shows the main experiment results. The best recall and F1 (0.60, 0.75) were obtained in the All-Variants/Without-US condition; the best accuracy (0.91) in the HT+TL-only conditions and PUOH-only/Without-US; and the best precision (0.83) in HT+TL-only/Without-US. Of the single-variant conditions, OHD-only performed best. Excluding the Unsure subclass improved results in almost every condition; the one partial exception was PUOH+OHD, where With-US gave marginally higher precision and accuracy. Confusion matrices (not shown for space) indicate that in the best-performing condition, true negatives were most common (290/494), followed by true positives (108), false negatives (61), and false positives (35); in the harder HT+TL-only/Without-US condition, the model predicted a positive label for only 16 of 372 instances (7 correct), indicating a strong majority-class bias.

The follow-up experiment, which fixed the sample size at 250 images per class, yielded more homogeneous results: OHD-only (both US conditions) and All-Variants/Without-US all reached F1 and accuracy of 0.76, while All-Variants/With-US performed worse (0.70). Relative to the main experiment, precision, recall, and F1 improved in nearly all conditions once sample-size was controlled for, while accuracy decreased slightly, consistent with accuracy's sensitivity to class imbalance; this indicates that the class-imbalance effect, not variation in form itself, was largely responsible for the poorer scores observed for smaller-subclass conditions in

	With-US		Without-US	
Condition	F1	Acc	F1	Acc
PUOH-only	0.63	0.90	0.71	**0.91**
OHD-only	0.68	0.84	0.74	0.85
HT+TL-only	0.50	**0.91**	0.59	**0.91**
PUOH+OHD	**0.74**	0.82	**0.75**	0.81
OHD+HT+TL	0.69	0.80	0.74	0.81
All-Variants	0.73	0.78	**0.75**	0.79

Table 3: Macro F1 and accuracy of the static model (main experiment). Precision and recall are omitted for space; the model achieved its highest recall (0.60) and precision (0.83) in All-Variants/Without-US and HT+TL-only/Without-US respectively.

	With-US		Without-US	
Condition	F1	Acc	F1	Acc
OHD-only	0.45	0.55	**0.53**	**0.59**
All-Variants	0.26	0.30	0.27	0.33

Table 4: Macro F1 and accuracy of the dynamic model. Precision and recall track F1 closely in all four conditions (within 0.10).

the main experiment.

6.2 Dynamic Model

Table 4 shows the dynamic model's results across its four conditions. OHD-only/Without-US performed best on every metric (precision, recall, F1 = 0.53; accuracy = 0.59); the All-Variants/With-US condition performed worst on recall, F1, and accuracy (0.34, 0.26, 0.30) despite a comparatively high precision (0.48). Confusion matrices show that in the two OHD-only conditions the negative class was most frequently (and largely correctly) predicted, whereas in the All-Variants conditions the no-gesture label dominated predictions for all classes: in All-Variants/With-US the positive label was predicted for only 16 of 167 true positive clips (10 correct), and in All-Variants/Without-US it was never predicted at all. This indicates that increased variation in the positive class pushed the model's PUOH-probability below the 0.9 classification threshold even for genuinely positive clips.

6.3 Individual Variation

Per-participant macro F1 was computed for the best static model (All-Variants/Without-US) and best dynamic model (OHD-only/Without-US) (Table 5). Scores varied substantially across participants: 0.36–1.00 for the static model and 0.32–0.84 for the dynamic model. For the static model, F1 correlated significantly and positively with a participant's number of test instances (Pearson's $r = 0.46$, $p = 0.04$), indicating sensitivity to per-

	Static		Dynamic	
Participant	F1	Sample	F1	Sample
1L	0.44	5	0.44	5
1R	0.71	38	0.52	27
2L	**0.38**	5	**0.84**	7
2R	**1.00**	4	0.46	9
3L	0.60	41	0.34	45
3R	0.75	33	**0.32**	19
4L	0.46	13	0.49	13
4R	0.42	11	0.76	14
5L	0.63	28	0.54	28
5R	0.42	18	0.66	15
6L	0.69	23	0.35	20
6R	0.79	40	0.57	31
7L	**0.85**	61	0.50	45
7R	0.75	24	0.50	16
8L	0.80	15	0.43	9
8R	0.81	34	0.58	32
9L	0.60	30	0.44	26
9R	0.80	33	**0.32**	34
10L	0.78	24	0.33	12
10R	0.36	14	0.61	10

Table 5: Macro F1 score and sample size in test set per participant for the All-Variants/Without-US condition of the static model and the OHD-only/Without-US condition of the dynamic model in the main experiment.

participant sample size; the one clear exception (a perfect F1 from only four test instances) is likely attributable to chance given the small sample. No significant correlation was found for the dynamic model ($r = -0.28$, $p = 0.23$).

7 Discussion

7.1 Interpretation

The static model's results correlate strongly with the relative size of the positive class: conditions with more positive instances (e.g., All-Variants) yield the best recall, precision, and F1, while conditions with fewer positive instances (e.g., HT+TL-only, PUOH-only) yield the highest raw accuracy, an artifact of accuracy's sensitivity to class imbalance rather than evidence of better classification. The follow-up experiment supports this: once sample size was controlled for, the model performed worst specifically in the condition combining maximal positive-class variation with the Unsure subclass (All-Variants/With-US), suggesting that variation in the positive class and the presence of near-miss negative instances compound one another, either because Unsure gestures resemble specific positive subclasses that only co-occur in this condition, or because greater positive-class variation broadens the model's implicit criterion for a positive label enough to admit Unsure instances.

For the dynamic model, the reverse pattern held:

performance was better with less positive-class variation (OHD-only) despite greater class imbalance, and markedly worse under All-Variants, where the model rarely predicted the positive label at all. A plausible explanation is architectural: the static model's features describe only hand shape, and the dynamic model's features span the full body (arms, shoulders, face), so that arm and shoulder movements distinguishing larger PUOH variants may be informative for the static model but a source of noise for the dynamic one.

Both models showed comparable sensitivity to individual variation in the participant-level analysis, though only the static model showed a significant relationship with sample size, suggesting the two models may be sensitive to different sources of inter-participant variation.

7.2 Error Analysis

A qualitative error analysis (samples of 10 images/videos per confusion-matrix category) surfaced three recurring patterns for both models. First, both models appear more sensitive to hand *shape* (open vs. closed) than to hand *orientation* (palm up vs. down): open hands with the palm facing down were repeatedly misclassified as positive PUOH instances, and this was the single most consistent source of false positives in both models. Second, incomplete visibility of the hand due to occlusion, camera angle, or video trimming to remove the other participant, degraded performance more for the dynamic model than the static model, and for the static model specifically, false negatives disproportionately involved gestures made with only one hand, consistent with the model receiving only single-hand landmarks as input. Third, for the dynamic model, a short hold phase (where retraction followed the stroke almost immediately) was associated with missed positive instances, suggesting the model relies substantially on the hold rather than the full gesture trajectory. Several instances in the negative-class samples that a human annotator would likely mislabel as positive without full video context were nonetheless correctly rejected by both models, indicating the models pick up on some cues that are not obvious to annotators working from static images alone, though this could also partly reflect the negative class's status as the training majority.

	Static	Dynamic
Better overall performance	X	
Requires little manual labor		X
Mainly limited by class imbalance	X	
Mainly limited by form variation		X
Input features	hand	full body
Max. hands used as input	1	2
Output classes	2	3

Table 6: Comparison of the static and dynamic models.

7.3 Comparison of Approaches

Table 6 summarizes the trade-offs between the two approaches. The static model achieved better overall performance and required under a minute to train and evaluate, versus roughly ten hours for the dynamic model, but required substantially more manual labor, since gesture holds had to be located and screenshotted by hand, whereas the dynamic model could use automatically extracted clips directly. The static model's single binary output is also simpler than the dynamic model's two-stage motion-then-gesture-type classification, though the latter's added complexity could in principle support classifying additional gesture types. Based on these results, the static model seems to be the more practical choice for naturalistic data (typically imbalanced and highly variable in form, as here) while the dynamic model may be better suited to more controlled, artificial recording settings. However, as is further discussed in Section 8, the observed results could also be due to a lack of training time and data for the dynamic model.

7.4 Connection to the Literature

That both models can recognize the PUOH gesture despite substantial within-class variation supports Cienki (2021)'s claim that its different forms share underlying features across the continuum. Consistent with this, both models' most reliable feature appears to be hand shape (an external-system feature; Bressem et al., 2013), both also struggle most with lower-arm rotation and supination, a feature central to Boutet et al. (2018)'s internal, joint-based system, lending some support to Cienki's argument that the internal system better captures what unifies these gesture forms. The results also speak to Lutgen et al. (2026)'s broader argument for treating variation as signal rather than noise in NLP data: multiple PUOH forms were found to co-occur naturally in Dutch conversation, gesture frequency varied substantially by participant (consistent with Chu et al., 2014), and, in the follow-up experiment,

choosing what to represent in the data (a fixed, balanced sample) outperformed simply maximizing data volume, directly supporting Lutgen et al.'s recommendation to base data-selection choices on desired representation rather than availability.

8 Conclusion

This paper investigated how gesture recognition models can recognize recurrent gestures in Dutch, focusing on the palm up open hand gesture. A small corpus of naturalistic Dutch dyadic conversation was collected and annotated, in which the PUOH gesture, previously documented in many spoken and signed languages but not, to this knowledge, in Dutch specifically, was found to occur frequently (567 instances) and in the range of forms described by Cienki (2021), addressing RQ1. Two recognition models, a static image-based model and a dynamic video-based model, were trained and compared across conditions varying in positive-class variation and class balance to address RQ2. The static model was found to perform better overall and to be less affected by increased form variation, and the dynamic model performed better only when positive-class variation was minimized; both models struggled most with distinguishing palm orientation from hand shape. These results indicate that automatic recognition of the PUOH gesture is possible, though a great deal of further work is needed to build higher-performing, more robust models, and that real-world gesture variation, an essential part of how individuals and groups use language to express themselves, should remain a central focus of that work.

Future Work. Both models could benefit from more training data, and from data recorded from multiple or varied camera angles, to determine how sensitive current results are to viewpoint. Comparing additional recognition architectures beyond the two considered here, and specifically studying how models handle palm orientation (e.g., by constructing data that isolates this variable), would help address the orientation-confusion errors observed in both models. For the static model specifically, incorporating arm and shoulder landmarks alongside hand landmarks, automating the currently manual hold-detection step, and controlling for per-participant sample size during training are promising directions. For the dynamic model, testing performance on artificially elicited (rather than naturalistic) PUOH productions would help isolate the effect of positive-class variation from other confounds. More broadly, extending this approach to other recurrent gestures, and to other languages, could make quantitative analysis of recurrent gestures considerably more efficient.

Limitations

The corpus is small (20 speakers, one regional variety of Dutch, a narrow age and education range) and drawn from a single elicitation task, limiting how far these findings generalize. The static model uses only one hand's landmarks as input, which likely contributes to false negatives on gestures produced with the non-tracked hand or with both hands together. The dynamic model was trained for only 5 of the 500 epochs used in the architecture it was adapted from, owing to time and compute constraints; it was still improving when training was stopped, so its reported performance likely understates what the architecture could achieve with full training. Video resolution was in some cases too low to clearly resolve hand landmarks, which may have affected the static model's input quality. Finally, because gesture subclass boundaries lie on a continuum, subclass labels (and the Unsure category in particular) involved some annotator judgment, which may not generalize across annotators.

Ethics Statement

This study was reviewed and approved by the Research Ethics Review Committee and Research Integrity of the host institution prior to data collection. All participants provided informed consent, including separate consent regarding the potential future use of anonymized data excerpts in publications or presentations. Recordings and derived materials are stored on institutional secure cloud storage; only the subset of the corpus covered by participants' consent will be shared with other researchers upon request.

Acknowledgments

This research was funded by the Department of Language, Literature and Communication (LLC) at VU Amsterdam School of Humanities. The authors would like to extend special thanks to the lab managers of the Social Robotics lab, who provided the necessary equipment, as well as the students who participated in this study.

References

Ahmad Sami Al-Shamayleh, Rodina Ahmad, Mohammad AM Abushariah, Khubaib Amjad Alam, and Nazean Jomhari. 2018. A systematic literature review on vision based gesture recognition techniques. *Multimedia Tools and Applications*, 77(21):28121–28184.

Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and social psychology bulletin*, 21(4):394–405.

Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse processes*, 15(4):469–489.

Dominique Boutet, Aliyah Morgenstern, and Alan Cienki. 2018. Looking ahead: Kinesiological analysis. *Aspectuality cross Languages Event construal in speech and gesture*, pages 143–159.

Jana Bressem and Cornelia Müller. 2014. A repertoire of german recurrent gestures with pragmatic functions. *Body–language–communication: An international handbook on multimodality in human interaction*, 2:1558–1574.

Jana Bressem et al. 2013. A linguistic perspective on the notation of form features in gestures.

Feng-Sheng Chen, Chih-Ming Fu, and Chung-Lin Huang. 2003. Hand gesture recognition using a real-time tracking method and hidden markov models. *Image and vision computing*, 21(8):745–758.

Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. 2019. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10(1):131–153.

Mingyuan Chu, Antje Meyer, Lucy Foulkes, and Sotaro Kita. 2014. Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, 143(2):694.

A Cienki, O Iriskhanova, A Agafonova, M Kiose, A Leonteva, A Petrov, O Prokofeva, and E Smirnova. Forthcoming. Recurrent gestures and verbal stance markers in russian: Viewing each in the light of the other. *Gesture*.

Alan Cienki. 2021. From the finger lift to the palm-up open hand when presenting a point: A methodological exploration of forms and functions. *Languages and Modalities*, 1(1):17–30.

Alan Cienki, Maíra Avelar, and Olga Iriskhanova. 2025. Action schemas can account for prototypical and non-prototypical palm-up open hand gestures. *Languages*, 10(11):282.

Kensy Cooperrider, Natasha Abner, and Susan Goldin-Meadow. 2018. The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, 3:23.

Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished language technology: The lack of (social) class in nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8675–8682.

Trevor Darrell and Alex Pentland. 1993. Space-time gestures. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340. IEEE.

Pierre Feyereisen and Jacques-Dominique De Lannoy. 1991. *Gestures and speech: Psychological investigations*. Cambridge University Press.

Jana M Iverson and Susan Goldin-Meadow. 1998. Why people gesture when they speak. *Nature*, 396(6708):228–228.

Jana M Iverson and Susan Goldin-Meadow. 2005. Gesture paves the way for language development. *Psychological science*, 16(5):367–371.

Annelies Jehoul, Geert Brône, and Kurt Feyaerts. 2017. The shrug as marker of obviousness: corpus evidence from dutch face-to-face conversations. *Linguistics Vanguard*, 3(s1):20160082.

Adam Kendon. 1995. Gestures as illocutionary and discourse structure markers in southern italian conversation. *Journal of pragmatics*, 23(3):247–279.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Minwoo Kim, Jaechan Cho, Seongjoo Lee, and Yunho Jung. 2019. Imu sensor-based hand gesture recognition for human-machine interfaces. *Sensors*, 19(18):3827.

Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167.

Ruslan Kurdyumov, Phillip Ho, and Justin Ng. 2011. Sign language classification using webcam images. *Comput. Therm. Sci*, 10:9029.

Silva H Ladewig. 2014. 118. Recurrent gestures. In *Volume 2*, pages 1558–1574. De Gruyter Mouton.

James Loeb. 1912. The Loeb classical library.

Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in nlp. *arXiv preprint arXiv:2603.24222*.

Max Planck Institute for Psycholinguistics, The Language Archive. 2026. ELAN (version 7.1) [computer software]. https://archive.mpi.nl/tla/elan. Accessed: 2026-07-06.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Joseph Albert Mosher. 1916. *The essentials of effective gesture, for students of public speaking*. Macmillan.

Cornelia Müller. 2004. Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256.

Cornelia Müller. 2010. Wie gesten bedeuten eine kognitiv-linguistische und sequenzanalytische perspektive. *Sprache und Literatur*, 41(1):37–68.

Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill, and Sedinha Tessendorf. 2013. *Body-Language-Communication. Volume 1*. Walter de Gruyter.

Yanghee Nam and KwangYun Wohn. 1996. Recognition of space-time hand-gestures using hidden markov model. In *Proceedings of the ACM symposium on virtual reality software and technology*, pages 51–58.

Juan Jesus Ojeda-Castelo, Maria de Las Mercedes Capobianco-Uriarte, Jose Antonio Piedra-Fernandez, and Rosa Ayala. 2022. A survey on intelligent gesture recognition techniques. *IEEE Access*, 10:87135–87156.

Edward Amherst Ott. 1902. *How to gesture*. Hinds & Noble.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv preprint arXiv:1608.07836*.

Wim Pouw. 2024. Envisionbox modules for social signal processing. https://github.com/WimPouw/envisionBOX_modulesWP. [Computer software]. Accessed: 2026-07-06.

Olga Prokofyeva, Maria Kiose, Anna Leonteva, and Evgeniya Smirnova. 2025. Intensity and its manifestations in speech and recurrent gestures in spontaneous dialogue. In *Proceedings of the International Conference "Dialogue*, volume 2025.

Alan Ramponi. 2024. Language varieties of italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Jürgen Erich Schmidt. 2010. Language and space: The linguistic dynamics approach. *Language and space. An international handbook of linguistic variation*, 1:201–225.

Keshav Sinha, Rashmi Kumari, Annu Priya, and Partha Paul. 2019. A computer vision-based gesture recognition using hidden markov model. In *Innovations in Soft Computing and Information Technology: Proceedings of ICEMIT 2017, Volume 3*, pages 55–67. Springer.

William C Stokoe Jr. 1960/2005. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37.

Jürgen Streeck. 2005. Pragmatic aspects of gesture. *International encyclopedia of languages and linguistics*, 5:71–76.

Jürgen Streeck. 2009. Gesturecraft.

Shigeyuki Tateno, Yiwei Zhu, and Fanxing Meng. 2019. Hand gesture recognition system for in-car device control based on infrared array sensor. In *2019 58th Annual conference of the society of instrument and control engineers of Japan (SICE)*, pages 701–706. IEEE.

Sedinha Teßendorf. 2014. Pragmatic and metaphoric–combining functional with cognitive approaches in the analysis of the "brushing aside gesture". *Body–language–communication: An international handbook on multimodality in human interaction*, 2:1540–1557.

Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, Mohamed K Shahin, and Basma Refaat. 2015. Sift-based arabic sign language recognition system. In *Afro-European Conference for Industrial Advancement: Proceedings of the First International Afro-European Conference for Industrial Advancement AECIA 2014*, pages 359–370. Springer.

Odlibek Tokhirov. 2023. How i built a hand gesture recognition model in python. Link. Accessed: 06-07-2026.

Pei Xu. 2017. A real-time hand gesture recognition and human-computer interaction system. *arXiv preprint arXiv:1704.07296*.

Bosco Yung. 2022. Nodding pigeon. https://github.com/bhky/nodding-pigeon. [Computer software]. Accessed: 2026-07-06.

Han Zou, Yuxun Zhou, Jianfei Yang, Hao Jiang, Lihua Xie, and Costas J Spanos. 2018. Wifi-enabled device-free gesture recognition for smart home automation. In *2018 IEEE 14th international conference on control and automation (ICCA)*, pages 476–481. IEEE.

The Hidden Cost of Personality: How Conditioning Methods Disrupt Gesture Generation in Small Language Models

Tova Radhe
Uppsala University
Sweden
tova.radhe.1246@student.uu.se

Alessio Galatolo
Uppsala University
Sweden
alessio.galatolo@it.uu.se

Abstract

Personality is a key factor in human–agent and human-robot interaction, influencing user perception, trust, and engagement. This work investigates methods for inducing Big Five personality traits in (open) language models, with a focus on smaller variants as more easily deployable in real-time agents and robotic systems. We compare prompt-based conditioning, parameter-efficient fine-tuning, and activation steering, focusing on their behaviour in multi-turn dialogue and embodied gesture generation. In particular, we investigate personality drift across turns and non-verbal behaviour generation through direct function calls.

We find that classifier-assigned trait alignment differs little across methods, with no consistent evidence of trait drift over multi-turn interaction. However, conditioning method substantially affects dialogue quality and robustness, with steering and fine-tuning introducing artefacts and instability. Interestingly, the more consequential differences emerge in the multimodal channel. Conditioning method, not intended personality, is the dominant driver of gesture behaviour, reshaping technical tool-use reliability, structural gesture diversity, and output-gesture congruence. Our findings suggest that personality conditioning should be evaluated jointly across verbal and non-verbal channels, and that method choice is itself a design decision shaping the expressiveness and reliability of embodied conversational agents.

1 Introduction

Personality in Human-Agent and Human-Robot Interaction (HRI) typically refers to behavioural patterns that lead users to attribute human-like traits to an agent or robot (Cerekovic et al., 2014; von der Pütten et al., 2010; Mou et al., 2020). Previous works show that user acceptance, trust and engagement are influenced by a robot's perceived personality (Whittaker et al., 2021; Goetz and Kiesler, 2002), from early hand-crafted behavioural designs (Lee et al., 2006; Tapus et al., 2008; Andrist et al., 2015) to more recent context-dependent approaches (Andriella et al., 2021). These traits are often inspired by psychological models of personality, with the Big Five framework (McCrae and Costa, 1987; Goldberg, 1990) being one of the most commonly used in these studies (Gittens, 2018; Robert et al., 2020). This framework describes personality across the five traits of Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism, where each trait is assigned either a binary (high/low) or continuos value score (0-1).

Recent developments in generative language models have introduced new possibilities for implementing personality in these systems. Large Language Models (LLMs) exhibit measurable and controllable personality traits (Serapio-García et al., 2025; Jiang et al., 2024; Lee et al., 2025a), most commonly induced through prompt-based conditioning (Jiang et al., 2023; Cho and Cheong, 2025). However, prompted personas are known to degrade in multi-turn interaction, producing repetitive dialogue or generic "as an AI assistant" artefacts (Bhandari et al., 2025b; Pinto-Bernal et al., 2025; Irfan et al., 2025), a problem that is exacerbated in Small Language Models (SLMs) (Wang et al., 2025), which are otherwise attractive for real-time agents or embedded, resource-constrained robotic deployment.

Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2022; Liu et al., 2022) and activation steering (Turner et al., 2024; Zou et al., 2025) have been proposed as alternative methods to personality conditioning by modifying the models internal representations rather than relying solely on system prompts (Li et al., 2025; Sun et al., 2025; Bhandari et al., 2025a). While these methods show promise in controlled evaluations, they are often assessed using self-report personality measures, which do not necessarily reflect how personality manifests in

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 160–174
October 29, 2026 ©2026 Association for Computational Linguistics

real interactive settings (Sikström et al., 2025), and additionally miss to capture changes in interaction quality.

Personality perception is also shaped by non-verbal, expressive behaviour, and recent work uses LLMs as high-level controllers that map dialogue content to non-verbal behaviour such as facial gestures (Mahadevan et al., 2024; Abbo et al., 2026; Antony et al., 2025). How underlying personality configurations influence this gesture generation, however, remains largely unexplored.

Our contributions are as follows:

- We provide a systematic comparison of prompting, PEFT, and activation steering for Big Five personality conditioning in small language models, evaluated under a shared protocol across multiple models and both dialogue and embodied settings.
- We show that classifier-assigned personality expression differs little across conditioning methods and shows no consistent drift over multi-turn dialogue, while dialogue quality diverges sharply, with steering and PEFT introducing substantially more artefacts and instability than prompting.
- We introduce a gesture-generation evaluation for the Furhat robot and find that conditioning method, not intended personality, is the dominant driver of gesture behaviour, degrading gesture diversity, output-gesture congruence, and tool-use reliability in smaller models, decoupled from effects on verbal personality expression.[1]

2 Related Work

2.1 Personality Conditioning with LLMs

Previous work shows that LLMs' personality is measurable and shapeable: applying validated psychometric instruments, Serapio-García et al. (2025) and Jiang et al. (2024) show that assigned Big Five personas are expressed consistently enough for human raters to identify them above chance, while Lee et al. (2025a) introduce a large benchmark confirming that this consistency is shaped substantially by alignment training. Prompt-based conditioning (Jiang et al., 2023; Cho and Cheong, 2025) remains the dominant method but is unstable in multi-turn or embodied settings, where personas drift, repeat, or collapse into generic responses (Atreja et al., 2025; Grabb, 2023; Almeida, 2025; Lee et al., 2025b; Bhandari et al., 2025b; Pinto-Bernal et al., 2025).

Alternative methods modify the model itself rather than relying solely on the prompt. Li et al. (2025) and Cui et al. (2024) use fine-tuning and preference optimisation on trait-grounded dialogue data to instil traits more stably than prompting. A separate line intervenes on internal activations at inference time without weight updates, steering personality-related directions in the residual stream (Sun et al., 2025; Bhandari et al., 2025a), extracting "persona vectors" for monitoring and control (Chen et al., 2025), or manipulating personality-related neurons directly (Deng et al., 2025). These methods are evaluated almost exclusively with self-report measures; trait directions have further been shown to be context-dependent (Frising and Balcells, 2026) and not fully geometrically independent, causing interference when multiple traits are steered jointly (Bhandari et al., 2026). Persona conditioning also affects content safety: assigning ChatGPT a persona can increase toxicity in generated text by up to a factor of six (Deshpande et al., 2023).

2.2 Personality Conditioning in HRI

Robot personality has been studied in HRI for two decades, predating the use of LLMs. Lee et al. (2006) showed that users can accurately perceive a robot's designed extraversion-introversion; subsequent work expressed and adapted this trait through proxemics and vocal delivery (Tapus et al., 2008), coordinated speech and gesture (Aly and Tapus, 2013), and gaze behaviour (Andrist et al., 2015), later extending to context-dependent, multi-trait personalities (Andriella et al., 2021). These largely hand-crafted, extraversion-focused designs are synthesised in existing reviews (Robert et al., 2020; Mou et al., 2020).

LLMs have begun to remove this scripting bottleneck by generating or modulating a robot's personality directly through dialogue, conditioning robots such as NAO (Zhang et al., 2025), NICO Becker et al. (2025), Furhat (Cherakara et al., 2023; Galatolo et al., 2023), and robotic arms (Nardelli et al., 2025). In parallel, gesture generation research has focused on mapping dialogue to gestures using prompting, without considering the effect of per-

[1] Note that our evaluation focuses on structural and technical gesture selection via tool calling, rather than human perceptual judgment of gesture naturalness.

sonality (Abbo et al., 2026; Antony et al., 2025; Nichols et al., 2026).

2.3 Literature Gaps

This leaves three gaps: comparison of conditioning methods in multi-turn interaction, their impact on dialogue quality, and their influence on gesture generation. To address these questions, this work investigates personality conditioning in SLMs across three dimensions: trait alignment in multi-turn dialogue, impact on dialogue quality and artefacts, and effects on gesture generation, with a focus on facial gestures. Prompt-based conditioning is compared with PEFT and activation steering, enabling a systematic evaluation on how different methods influence both verbal and non-verbal behaviour in embodied agents.

3 Methodology

We compare three personality-conditioning methods, prompting, PEFT, and activation steering, under a shared protocol, so that observed differences reflect the conditioning method rather than differences in data or evaluation. All three are conditioned on the same BIG5-CHAT subset (Li et al., 2025), consisting of 10 000 dialogue input/output pairs based on human-grounding data for each Big Five trait and level. Each conditioned model is placed in 100 simulated ten-turn dialogues driven by a fixed interlocutor and a trait classifier scores every response, yielding both an average measure of trait expression and its trajectory across turns. The same models are then evaluated in an embodied setting, where they select facial gestures via tool calling for the social robot Furhat (Al Moubayed et al., 2012), testing whether conditioning effects extend beyond text. We chose Furhat as the platform as its facial features could extend both to robotics settings as well as virtual agents.

We evaluate across three models spanning two size regimes and two model families: Qwen3-4B-Instruct-2507 (Yang et al., 2025), Qwen3-1.7B (Yang et al., 2025), and Ministral-3-3B-Instruct-2512 (Liu et al., 2026).

3.1 Inducing Personality

Prompting. We adopt the instruction- and demonstration-based (few-shot learning) prompts from Li et al. (2025). While the authors report a slight advantage for instruction-based prompting, we observed it induced several meta-artefacts in the generated dialogue during initial trial runs. Specifically, the model would explicitly verbalise its assigned trait (e.g., explicitly stating it had low agreeableness), or even attributing the trait to the user, disrupting natural conversation flow. Because demonstration-based prompting yielded more natural dialogue during in comparison, we selected it as our primary prompting method. We report in Appendix A.1 the final prompt used.

PEFT. We fine-tune each model using QLoRA (Dettmers et al., 2023) via TRL (von Werra et al., 2020). Using the BIG5-CHAT dataset (Li et al., 2025), we train a separate adapter with a subset for each trait and level, resulting in 10 fine-tuned personalities in total for each model. Full training settings for each model are in Appendix A.2.

Steering. Using the EasySteer framework (Xu et al., 2025) with inspiration from Bo et al.'s replication (Bo et al., 2025), five trait vectors for the Big Five personality traits were created, utilising 250 randomised high/low trait pairs from the BIG5-CHAT dataset as the contrast pairs.

Injection scale and the set of steered layers interact strongly with model depth, so we select them per model by a comparative qualitative assessment for a controlled sample. See Appendix A.3 for details on selection. This yields a scale of ± 1.8 applied across all 36 layers for Qwen3-4B, ± 4 across all 28 layers for Qwen3-1.7B, and 0.4 applied to layers 10–15 for Ministral.

3.2 Dialogue Generation and Evaluation

Simulated dialogues. To probe personality expression in dialogue, we generate multi-turn conversations between the personality-conditioned models and a fixed conversational agent. To test the hypothesised personality drift over dialogue turns, we evaluate the conversation over 10 turns, where each turn includes both a 'simulated' user input and a corresponding model response. We utilise Qwen3-4B-Instruct-2507 as the fixed interlocutor for all conditioned models. To ensure diversity of interaction contexts, we sample 100 conversation topics from the Langame conversation-starters dataset[2], where for each conversation, a topic was injected into the interlocutor's prompt, guiding the interaction while allowing for natural dialogue progression. See Appendix A.1 for full prompt details. The personality-conditioned models did not receive

[2] HuggingFace: Langame/conversation-starters

any specific prompt (excluding the prompt conditioning variant). Dialogues are generated for every combination of conditioning method (prompting, PEFT, steering), personality trait (Big Five), trait level (high/low) and for the base models (the unmodified versions of selected language models).

Measuring Big Five traits from text. While it is common in the literature to evaluate LLMs' personality by directly assessing it using standard human questionnaires (Miotto et al., 2022; Jiang et al., 2023; Serapio-García et al., 2025), this inherently presents methodological flaws, as also outlined by Serapio-García et al. (2025), e.g., inconsistencies between reported traits and traits exhibited in texts (e.g., it has been shown that LLMs 'know' when they are being tested and may exhibit the expected 'desired' behaviour, Needham et al. (2025)). For this reason, we chose to measure personality expression by analysing the model output in real-world settings and assessing its personality using the `Personality_LM` classifier (Wang and Sun, 2024), which assigns scores corresponding to the Big Five traits, scores ranging from 0-1 where 1 indicates higher trait alignment. Accuracy and F1 scores for Personality_LM reach 80% and 79% respectively (Wang and Sun, 2024), as reported by the authors, resulting in a rather reliable tool. For each conversation, personality expression is summarised using the mean of the target trait across all turns. In addition, we compute the mean for each individual turn, enabling assessment of personality drift throughout the dialogue.

Dialogue quality and artefacts. The dialogue outputs were analysed for common dialogue artefacts that may negatively affect interaction quality. Namely: emoji overuse, empty responses, and AI meta responses. The latter refers to outputs where the model explicitly describes itself or its limitations (e.g., "As an AI model..."). These instances were identified through keyword-based text search. Empty responses were detected by identifying missing or blank entries in the generated output data (e.g., empty fields in the recorded CSV files). Emoji overuse was assessed by comparing the average frequency of emoji usage in the base model outputs to those generated under different personality configurations.

3.3 Gesture Generation

In humans, personality affects gesture production (Hostetter and Potthoff, 2012; Lopez-Ozieblo, 2024; Argyle, 2013; Riggio et al., 1990). For virtual agents or social robots, reinforcing personality expression through co-speech gestures has been shown to help with *perception* of the intended personality (Neff et al., 2010; Sonlu et al., 2024; Smith and Neff, 2017; Niu et al., 2022). Among HRI practitioners, it has also become common to generate co-speech gestures utilising tool and function calling (Mahadevan et al., 2024; Arenas et al., 2024). Our goal is thus to evaluate how gesture generation differs across personality configurations and conditioning methods when tasked with selecting gestures via tool calling in response to an input. For each input, the model generates a structured JSON schema corresponding to a gesture from a predefined set representing the Furhat gesture library, see Appendix A.4 for the list of pre-defined Furhat gestures. In addition to gesture type, the model is also tasked to specify parameters for duration and intensity, sampled from discrete sets $\{0.7, 1, 1.5\}$ and $\{0.5, 1, 2\}$, respectively.

If the tool call is unsuccessful, either because no call is initiated by the model or the output does not conform to the required format, the result is defaulted to *No gesture*. Three experimental setups were used to evaluate gesture generation behaviour:

- **Controlled input**: Each model was given the same set of 10 textual inputs, enabling a controlled comparison of how different personality configurations influence gesture selection under identical conditions.
- **Output-gesture congruence**: Models were explicitly instructed to perform a specific gesture (e.g., "I am performing gesture {gesture}"), evaluating whether the generated gesture matches the intended output. This setup measures the congruence between verbal specification and produced behaviour.
- **Multi-turn conversation analysis**: Gesture generation was analysed across the previously generated conversations to assess multi-turn behaviour and how personality configurations influence gesture usage in dynamic interactions.

4 Results

We organise results around three key questions: whether conditioning shifts the target trait and persists across turns (§4.1), its effect on dialogue

quality (§4.2), and its effect on gesture generation(§4.3).

4.1 Does Conditioning Shift the Target Trait?

Trait alignment. Table 1 summarises the difference in classifier-assigned personality scores between high and low personality conditions across models and conditioning methods. Across all configurations, trait separation remains extremely limited, with Mean Δ values ranging between 0.0003 and 0.0022. This indicates that neither prompting, PEFT, nor steering produced substantial differences in measured personality expression. Steering generally yielded the largest trait differences, although the observed effects remain small in magnitude. Negative Δ values indicate cases where classifier scores moved in the opposite direction of the intended manipulation. However, because both the positive and negative effects were consistently close to zero, these reversals are better interpreted as evidence of weak personality induction rather than systematic counter-conditioning

Trait drift over time. Table 1 also reports the average drift in trait separation over the course of a conversation. Drift values remained close to zero across all models and conditioning methods, indicating that the already limited differences between high and low trait conditions neither consistently increased nor decreased over time. Contrary to our initial hypothesis, we therefore find little evidence that personality conditioning effects systematically degrade during multi-turn interaction.

4.2 Effects on Output Quality

Table 2 shows that conversational artefacts are relatively infrequent at the conversation level, but differ across models. The Ministral model exhibits empty responses in 12.4% of the conversations, meaning that at least one turn in these conversations is empty. No such failures are observed for the Qwen models.

On the other hand, AI-meta responses occur only in the Qwen models, appearing at least once in 2.3% of conversations for Qwen3-1.7B, 2.6% for Qwen3-4B.

Importantly, these values reflect whether an artefact occurs *at least once* within a conversation, rather than the total number of affected turns. As such, they capture the likelihood of encountering a problematic multi-turn interaction rather than frequency of errors within a single conversation.

Table 3 further highlights the configurations most associated with which conversational artefact. Across all models, steering dominates the highest frequency of compromised conversations.

For Ministral, empty responses are observed across all trait dimensions and levels, indicating that the issue is not limited to a specific personality configuration. However, certain conditions show higher concentrations, particularly *low neuroticism* and *low agreeableness*, suggesting that specific steering directions may negatively affect the model's conversational stability.

In Qwen3-1.7B, AI meta responses are highly concentrated in *low neuroticism* (52.2%) and *high openness* (43.4%), suggesting a strong dependence on specific trait manipulations. Qwen3-4B show in comparison a more diffuse pattern with lower peak values (19.2%) distributed across multiple traits.

We identified emoji overuse as a recurring artefact in both of the Qwen models, particularly under high extraversion steering. Analysis of the top 10 conversations with the largest deviation in emoji usage relative to their base model counterpart shows a strong concentration of emoji inflation with these configurations. In contrast, Ministral does not exhibit a consistent trend across traits. Emoji inflation occurs in multiple traits and is associated with the prompt-based method rather than with steering.

Overall, these results indicate that steering elicits the highest frequency of unwanted conversational artefacts across models, but the manifestation differs. Ministral is prone to structural failure with empty outputs, whereas Qwen models exhibit stylistic artefacts in the form of meta-responses. The results also suggest that artefact occurrence is both method-dependent and trait-sensitive rather than uniformly distributed.

4.3 Gesture Control and Selection

Selection and congruence. Gesture selection was compared across models and conditioning methods using identical input prompts. The results show that gesture distributions vary depending on the applied method, indicating that gesture selection is not solely determined by the input text but is also influenced by the personality conditioning approach.

Figure 1 summarises output-gesture congruence across models and personality conditioning methods. Congruence is measured as the proportion of generated gestures that match the requested target gesture in a controlled gesture-following task. All baseline models achieved perfect congruence,

Table 1: Trait separation between high and low personality conditions. Δ denotes the difference in classifier-assigned scores between high and low trait configurations. Mean |Δ| reports the average absolute high–low separation across all five personality traits. Avg Drift indicates change in trait separation from the first to the final dialogue turn. Positive Δ values indicate higher scores in the intended high-trait condition, whereas negative values indicate a reversal of the intended direction. Values near zero indicate limited personality differentiation and little evidence of trait drift.

Model	Method	ΔO	ΔC	ΔE	ΔA	ΔN	|Δ|Mean	Avg Drift
Ministral	Prompt	-0.0008	-0.0001	-0.0000	-0.0014	0.0002	0.0005	-0.0001
	PEFT	-0.0012	0.0005	-0.0013	0.0003	0.0003	0.0007	0.0006
	Steering	-0.0007	-0.0047	-0.0017	0.0007	-0.0004	0.0017	0.0001
Qwen3-1.7B	Prompt	-0.0002	0.0004	-0.0004	-0.0006	0.0001	0.0003	0.0003
	PEFT	-0.0006	0.0004	-0.0006	0.0002	0.0006	0.0005	0.0009
	Steering	-0.0001	-0.0034	-0.0017	-0.0021	0.0011	0.0017	0.0010
Qwen3-4B	Prompt	-0.0022	-0.0010	0.0008	0.0001	0.0007	0.0009	0.0010
	PEFT	-0.0002	-0.0002	-0.0004	-0.0004	0.0003	0.0003	-0.0001
	Steering	-0.0010	-0.0056	0.0026	0.0012	0.0005	0.0022	0.0003

Table 2: Proportion of conversations containing conversational artefacts across models. Percentages indicate the share of conversations in which the artefact occurs at least once.

Metric	**Ministral**	**Qwen3-1.7B**	**Qwen3-4B**
AI meta-responses	0%	2.3%	2.6%
Empty responses	12.4%	0%	0%

indicating that any reduction in performance is attributable to the personality conditioning methods rather than the underlying models. Qwen3-4B maintained near-perfect congruence across all methods, demonstrating strong robustness to personality conditioning. In contrast, Qwen3-1.7B and Ministral were substantially more sensitive to conditioning. Prompt-based conditioning generally preserved high congruence, whereas both PEFT and steering caused considerable degradation in gesture-following performance.

Overall, the results suggest that personality conditioning *can* interfere with multimodal behaviour generation, particularly for smaller models.

Gestures in multi-turn conversation. Table 4 summarises gesture generation during multi-turn conversations in terms of failure rate, hallucination rate, and gesture diversity. Failure rate measures the proportion of responses resulting in *No gesture*, hallucination rate captures the generation of gestures outside the predefined gesture inventory, and gesture diversity reports the number of unique valid gestures generated.

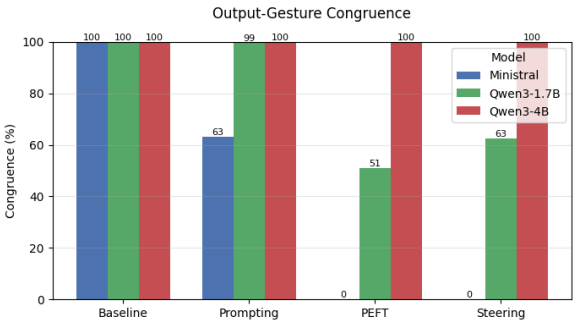


Figure 1: Output-gesture congruence across models and personality conditioning methods. Congruence is measured as the percentage of generated gestures that match the requested target gesture in a controlled gesture-following task. Qwen3-4B maintains near-perfect congruence across all methods, while PEFT and steering substantially reduce gesture-following accuracy for smaller models.

Table 3: Most artefact-prone configurations. Percentages represent the proportion of artefact-containing conversations within each model that are associated with the specified trait and method.

Model	**Artifact**	**Method**	**Trait (Level)**	**Percentage**
Ministral	Empty response	Steering	Agreeableness (Low)	21.8%
Ministral	Empty response	Steering	Neuroticism (Low)	21.0%
Qwen3-1.7B	AI meta-response	Steering	Neuroticism (Low)	52.2%
Qwen3-1.7B	AI meta-response	Steering	Openness (High)	43.4%
Qwen3-4B	AI meta-response	Steering	Conscientiousness (High)	19.2%
Qwen3-4B	AI meta-response	Steering	Extraversion (Low)	19.2%

Table 4: Gesture generation characteristics during multi-turn conversations. Failure Rate reports the proportion of responses resulting in *No gesture*, Hallucination Rate the proportion of generated gestures outside the predefined gesture inventory, and Gesture Diversity the number of unique valid gestures generated (excluding *No gesture*). In **bold**, the best performing method excluding the baseline.

Model	Method	Failure Rate (%)	Hallucination Rate (%)	Gesture Diversity
Ministral	*Baseline*	*15.20*	*0.00*	*8*
	PEFT	100.00	**0.00**	0
	Prompting	**36.50**	**0.00**	**14**
	Steering	100.00	**0.00**	0
Qwen3-1.7B	*Baseline*	*45.80*	*0.00*	*11*
	PEFT	99.96	**0.00**	2
	Prompting	**46.48**	**0.00**	10
	Steering	76.88	0.07	**15**
Qwen3-4B	*Baseline*	*3.80*	*0.00*	*6*
	PEFT	**22.05**	0.10	12
	Prompting	23.91	**0.06**	7
	Steering	26.07	0.67	**16**

Diversity and hallucinations. Across models, personality conditioning had a stronger effect on gesture behaviour than on measured personality expression. For Ministral and Qwen3-1.7B, both PEFT and Steering substantially increased failure rates, and both methods completely suppressed gesture generation for Ministral, producing *No Gesture* in all responses. Qwen3-1.7B exhibited a similar pattern, with failure rates approaching 100% under PEFT and exceeding 75% under steering. Inspection of failure modes showed that failures were predominantly caused by models not performing tool calls rather than malformed tool-call executions. Steering additionally introduced a larger proportion of formatting-related errors, suggesting that conditioning methods can affect both gesture generation and tool-use reliability. Qwen3-4B was considerably more robust. While all conditioning methods increased failure rates relative to the baseline, the model maintained expressive behaviour across conditions, producing between 7 and 16 unique gestures. Steering produced the highest gesture diversity, but was also associated with the highest hallucination rate. No hallucinated gestures were observed for Ministral, and only negligible rates were found for Qwen3-1.7B.

Overall, the results indicate that personality conditioning methods substantially influence multimodal behaviour generation. In smaller models, PEFT and steering frequently suppress gesture production and reduce interaction reliability, whereas Qwen3-4B largely preserves expressive behaviour while remaining susceptible to occasional hallucinated outputs.

5 Implications for Multimodal Interaction

The results suggest that personality conditioning methods impact the technical execution of non-verbal generation more strongly than they alter measured personality expression. Across models and conditioning approaches, classifier-assigned personality differences remained minimal and showed little evidence of drift over time. In contrast, the same conditioning methods produced substantial changes in gesture generation, tool-use reliability, and multimodal output quality.

Our results suggest that personality conditioning acts not only as a modifier of linguistic style, but also as an intervention on multimodal behaviour generation. While all methods produced only minimal differences in classifier-assigned personality scores, they substantially affected gesture generation, gesture diversity, output-gesture congruence, and tool-use reliability.

However, these effects were highly dependent on both conditioning method and the underlying model. In Ministral and Qwen3-1.7B, both PEFT and steering frequently suppressed gesture generation, resulting in high failure rates and reduced interaction reliability. Prompt-based conditioning generally preserved the behaviour of the base model more effectively, although failure rates still increased relative to baseline. In contrast, Qwen3-4B remained comparatively robust across conditioning methods, maintaining low failure rates and diverse gesture selection. Interestingly, steering increased gesture diversity in Qwen3-4B, but was also associated with a higher rate of hallucinated gestures.

Taken together, these findings suggest that evaluating personality conditioning through text-based personality assessments alone provides an incomplete picture. Models that exhibit similar personality scores may differ substantially in their non-verbal behaviour and multimodal interaction patterns. This highlights the importance of evaluating personality conditioning in settings that jointly consider verbal and non-verbal outputs. Overall, the choice of personality conditioning method should be considered not only as a mechanism for controlling dialogue style, but also as a design decision affecting the robustness, expressiveness, and reliability of multimodal interactive systems.

6 Conclusion

This work investigated the use of personality conditioning methods in small language models, focusing on their behaviour in multi-turn dialogue and gesture generation for embodied and virtual agents. We evaluated Prompt-based conditioning, PEFT, and activation steering across multiple models and all traits from the Big Five framework. Results show that differences in classifier-assigned personality scores between conditioning methods remain weak in multi-turn dialogue, with little to no evidence of drift throughout the conversation. However, differing conditioning methods have a substantial impact on dialogue quality. Steering and PEFT introduce instability, including increased artefacts and failure in adhering to structured outputs, while the prompt-based method remained more robust and in line with the base model's capabilities.

Personality conditioning additionally substantially influenced multimodal behaviour generation. Conditioning methods affected gesture diversity, gesture-output congruence, hallucinated gestures, and tool-use reliability. These effects were often considerably larger than the observed differences in personality expression itself. In particular, PEFT and steering frequently degraded multimodal performance in smaller models, while prompt-based conditioning generally preserved the behaviour of the underlying model.

Taken together, these results suggest that personality conditioning should not be viewed solely as a mechanism for controlling dialogue style. Even when conditioning methods produce little measurable change in personality expression, they can substantially alter non-verbal behaviours, tool-use, and the overall quality of multimodal interaction. As multimodal conversational agents increasingly integrate language with gesture, facial expression, and structured actions, personality conditioning should be viewed not only as a mechanism for shaping personality expression in dialogue, but also as an intervention that shapes the generation of multimodal behaviours and affects the reliability of interaction.

Future work should investigate conditioning methods that provide stronger behavioural control while maintaining reliable multimodal performance, as well as evaluation frameworks that jointly assess verbal and non-verbal aspects of interaction.

6.1 Limitations

Our study has some limitations that should be considered when interpreting the results. The comparison is restricted to three openly available small/mid-sized models, chosen for their relevance to embedded robotic deployment or real-time virtual assistants, rather than as a representative sample of the broader model landscape, so findings may not generalise to larger or closed-source systems. Personality expression was assessed using a single automatic classifier; although it reports high accuracy in prior work, we did not validate it against human ratings in our setting, and the near-zero trait separation we observe could partly reflect classifier insensitivity rather than an absence of behavioural change. Gesture correctness was evaluated through structural tool-call success against a predefined gesture inventory rather than perceptual appropriateness judged by human observers, and more generally, none of our findings are corroborated by human evaluation, which is particularly relevant given that personality perception in HRI is inherently human-centred. Finally, while prompting emerged as the most robust method, PEFT and steering require trait-specific training or vector extraction for every model and trait/level combination, incurring non-trivial computational and engineering overhead that is worth weighing against their limited gains in trait alignment.

Acknowledgements

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis, C3SE (Chalmers) partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Giulio Antonio Abbo, Ruben Janssens, Seppe Van De Vreken, and Tony Belpaeme. 2026. Expressive Furhat: Generating Real-Time Facial Expressions for Human-Robot Dialogue with LLMs. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pages 456–460, Edinburgh Scotland UK. ACM.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In *Cognitive Behavioural Systems*, volume 7403, pages 114–130. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Jothan Almeida. 2025. Prompt Engineering: A Comparative Study of Prompting Techniques in AI Language Models. In *2025 IEEE Integrated STEM Education Conference (ISEC)*, pages 1–4, Princeton, NJ, USA. IEEE.

Amir Aly and Adriana Tapus. 2013. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Antonio Andriella, Henrique Siqueira, Daniel Fu, Sven Magg, Pablo Barros, Stefan Wermter, Carme Torras, and Guillem Alenya. 2021. Do i have a personality? endowing care robots with context-dependent personality traits. *International Journal of Social Robotics*, 13:2081–2102.

Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: Matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*.

Victor Nikhil Antony, Maia Stiber, and Chien-Ming Huang. 2025. Xpress: A System for Dynamic, Context-Aware Robot Facial Expressions Using Language Models. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 958–967, Melbourne, Australia. IEEE.

Montserrat Gonzalez Arenas, Ted Xiao, Sumeet Singh, Vidhi Jain, Allen Ren, Quan Vuong, Jake Varley, Alexander Herzog, Isabel Leal, Sean Kirmani, Mario Prats, Dorsa Sadigh, Vikas Sindhwani, Kanishka Rao, Jacky Liang, and Andy Zeng. 2024. How to prompt your robot: A promptbook for manipulation skills with code as policies. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4340–4348.

Michael Argyle. 2013. *Bodily communication*. Routledge.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. What's in a Prompt?: A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:122–145.

Dennis Becker, Kyra Ahrens, Connor Gäde, Erik Strahl, and Stefan Wermter. 2025. Robots with attitudes: Influence of llm-driven robot personalities on motivation and performance. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI '25, page 40–49, New York, NY, USA. Association for Computing Machinery.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish

Nasim. 2025a. Activation-Space Personality Steering: Hybrid Layer Selection for Stable Trait Control in LLMs. *arXiv preprint*. Version Number: 1.

Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025b. Can LLM Agents Maintain a Persona in Discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29201–29217, Suzhou, China. Association for Computational Linguistics.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. 2026. Do personality traits interfere? geometric limitations of steering in large language models. *Preprint*, arXiv:2602.15847.

Jessica Y. Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2025. Steerable Chatbots: Exploring Personalization Control Interfaces via LLM Activation Steering. *arXiv preprint*. Version Number: 2.

Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2014. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. In *Human Behavior Understanding*, pages 1–15, Cham. Springer International Publishing.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, and Oliver Lemon. 2023. FurChat: An embodied conversational agent using LLMs, combining open and closed-domain dialogue with facial expressions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 588–592, Prague, Czechia. Association for Computational Linguistics.

Gunhee Cho and Yun-Gyung Cheong. 2025. Scaling personality control in llms with big five scaler prompts. *Preprint*, arXiv:2508.06149.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. 2024. Machine mindset: An mbti exploration of large language models. *Preprint*, arXiv:2312.12999.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao yang, Xin Zhao, and Ji-Rong Wen. 2025. Neuron based personality trait induction in large language models. In *International Conference on Learning Representations*, volume 2025, pages 85059–85083.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Michel Frising and Daniel Balcells. 2026. Linear personality probing and steering in llms: A big five study. *Preprint*, arXiv:2512.17639.

Alessio Galatolo, Iolanda Leite, and Katie Winkle. 2023. Personality-adapted language generation for social robots. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1800–1807.

Curtis L. Gittens. 2018. *A Psychologically-Realistic Personality Model for Virtual Agents*, pages 81–99. Springer International Publishing, Cham.

Jennifer Goetz and Sara Kiesler. 2002. Cooperation with a robotic assistant. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, pages 578–579, Minneapolis Minnesota USA. ACM.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Declan Grabb. 2023. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6:20–20.

Autumn B Hostetter and Andrea L Potthoff. 2012. Effects of personality and social situation on representational gesture production. *Gesture*, 12(1):62–83.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Bahar Irfan, Sanna Kuoppamäki, Aida Hosseini, and Gabriel Skantze. 2025. Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults. *Autonomous Robots*, 49(1):9.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can robots manifest personality? an empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication*, 56(4):754–772.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025a. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025b. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 100 others. 2026. Ministral 3. *arXiv preprint*. Version Number: 1.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Renia Lopez-Ozieblo. 2024. Is personality reflected in the gestures of second language speakers? *Frontiers in psychology*, 15:1463063.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative Expressive Robot Behaviors using Large Language Models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 482–491, Boulder CO USA. ACM.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Mou, Changqian Shi, Tianyu Shen, and Kun Xu. 2020. A Systematic Review of the Personality of Robot: Mapping Its Conceptualization, Operationalization, Contextualization and Effects. *https://doi.org/10.1080/10447318.2019.1663008*, 36(6):591–605.

Alice Nardelli, Lorenzo Landolfi, Dario Pasquali, Antonio Sgorbissa, Francesco Rea, and Carmine Recchiuto. 2025. Toward a universal concept of artificial personality: Implementing robotic personality in a kinova arm. *Preprint*, arXiv:2501.06867.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. Large language models often know when they are being evaluated. *Preprint*, arXiv:2505.23836.

Michael Neff, Yingying Wang, Rob Abbott, and Marilyn Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*, pages 222–235. Springer.

Eric Nichols, Miguel Mejia Tobar, and Randy Gomez. 2026. Holistic LLM-Based Expressive Behavior Generation for Robots. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pages 809–813, Edinburgh Scotland UK. ACM.

Jin Niu, Chih-Fu Wu, Xiao Dou, and Kai-Chieh Lin. 2022. Designing gestures of robots in specific fields for different perceived personality traits. *Frontiers in psychology*, 13:876972.

Maria Pinto-Bernal, Matthijs Biondina, and Tony Belpaeme. 2025. Designing Social Robots with LLMs for Engaging Human Interaction. *Applied Sciences*, 15(11):6377.

Ronald E Riggio, Richard Lippa, and Charles Salinas. 1990. The display of personality in expressive movement. *Journal of Research in Personality*, 24(1):16–31.

Lionel P. Robert, Jr, Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, and Qiaoning Zhang. 2020. A Review of Personality in Human–Robot Interactions. *Foundations and Trends® in Information Systems*, 4(2):107–212. ISBN: 9781680836165.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1954–1968.

Sverker Sikström, Ieva Valavičiūtė, and Petri Kajonius. 2025. Personality in just a few words: Assessment using natural language processing. *Personality and Individual Differences*, 238:113078.

Harrison Jesse Smith and Michael Neff. 2017. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Sinan Sonlu, Halil Özgür Demir, and Uğur Güdükbay. 2024. Personality expression using co-speech gesture. *ACM Trans. Appl. Percept.*, 22(2).

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality Vector: Modulating Personality of Large Language Models by Model Merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24667–24688, Suzhou, China. Association for Computational Linguistics.

Adriana Tapus, Cristian Tăpuş, and Maja J. Matarić. 2008. User–robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics*, 1:169–183.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Astrid M. von der Pütten, Nicole C. Krämer, and Jonathan Gratch. 2010. How our personality shapes our interactions with virtual characters - implications for research and development. In *Intelligent Virtual Agents*, pages 208–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Rong Wang and Kun Sun. 2024. Continuous output personality detection models via mixed strategy training. *ArXiv*.

Yixiao Wang, Homa Fashandi, and Kevin Ferreira. 2025. Improving Personality Alignment in Role-Playing Dialogue Agents for Small LLMs. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rome, Italy. IEEE.

Steve Whittaker, Yvonne Rogers, Elena Petrovskaya, and Hongbin Zhuang. 2021. Designing Personas for Expressive Robots. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(1).

Haolei Xu, Xinyu Mei, Yuchen Yan, Rui Zhou, Wenqi Zhang, Weiming Lu, Yueting Zhuang, and Yongliang Shen. 2025. Easysteer: A unified framework for high-performance and extensible llm steering. *arXiv preprint arXiv:2509.25175*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. Version Number: 1.

Alex Wuqi Zhang, Clark Kovacs, Liberto De Pablo, Justin Zhang, Maggie Bai, Sooyeon Jeong, and Sarah Sebo. 2025. Exploring Robot Personality Traits and Their Influence on User Affect and Experience. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 968–977, Melbourne, Australia. IEEE.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Appendix

A.1 Prompts

The following system-prompt was used for demonstration-based method:

`"You are helpfull assistant with {level} {trait}. Here are 10 examples of how people like you have responded in different situations. Pay attention to how they approach communication and problem-solving." + trait_examples_string`.

10 random examples from the BIG5-CHAT dataset with the desired trait and level were extracted and added as the variable `trait_examples_string`. Inspired by Li et al. (Li et al., 2025)

The following system prompt were used for the other methodologies, including base:

`"You are helpfull assistant."`

The following system prompts were used during gesture generation along with above system prompt:

"Select an appropriate facial gesture for the assistant's last message. You must call on perform_furhat_gesture()"

The following prompt was given to the language model performing as a user during conversations:

You are simulating a **human user** in a natural conversation with an AI assistant. Write like a person would (concise, 1-3 sentences that follow-up the AI assistants answer). Be curious and practical. No roleplay disclaimers, no meta-comments. Avoid repeating the assistant. Do not reveal you are an AI. Output ONLY the user's message, nothing else. Maintain this profile. You will be given a topic to discuss. Please generate a question in regards to the topic and stay on course.

The following output was hardcoded as the model's response to test output-gesture congruence:

I am performing this gesture: gesture

A.2 Hyperparameters

The following shows configurations for the fine-tuning across models.

Listing 1: PEFT configurations for the different models

```
#Qwen3-4B-Instruct-2507

peft_config = LoraConfig(
    r=32,
    lora_alpha=32,
    target_modules=['down_proj','
        o_proj', 'k_proj','q_proj'
        ,'gate_proj', 'up_proj','
        v_proj'],
)

base_config = {
    "num_train_epochs": 2,
    "per_device_train_batch_size"
        : 2,
    "gradient_accumulation_steps"
        : 8,
    "warmup_steps": 5,
    "learning_rate": 5e-5,
    "optim": "adamw_8bit",
    "max_length": None,
    "logging_steps": 1,
    "report_to": "trackio",
    "push_to_hub": True,
}

#Qwen3-1.7B
peft_config = LoraConfig(
    r=16,
    lora_alpha=16,
    target_modules=['down_proj','
        o_proj',
    'k_proj','q_proj','gate_proj'
        ,
    'up_proj','v_proj'],
)

base_config = {
    "num_train_epochs": 3,
    "per_device_train_batch_size"
        : 2,
    "gradient_accumulation_steps"
        : 8,
    "warmup_steps": 5,
    "learning_rate": 1e-4,
    "optim": "adamw_8bit",
    "max_length": None,
    "logging_steps": 1,
    "report_to": "trackio",
    "push_to_hub": True,
}

#ministral-3-3B-Instruct-2512
peft_config = LoraConfig(
    r=16,
    lora_alpha=16,
    lora_dropout=0.05,
    target_modules=['down_proj','
        o_proj', 'k_proj','q_proj'
        ,'gate_proj', 'up_proj','
        v_proj'],
)

base_config = {
    "num_train_epochs": 3,
    "per_device_train_batch_size"
        : 2,
    "gradient_accumulation_steps"
        : 8,
    "warmup_steps": 5,
    "learning_rate": 1e-4,
    "optim": "adamw_8bit",
    "max_length": None,
    "logging_steps": 1,
    "report_to": "trackio",
    "push_to_hub": True,
```

}

A.3 Steering Configuration Assessment

Qualitative analysis was conducted to determine appropriate vector configurations and layer application strategies for each model.

As shown in Figure 5, Ministral exhibited high sensitivity to layer selection, with certain layer ranges causing unintended behavior such as language switching to French. Additionally, a trade-off was observed between dialogue quality and strength of personality expression when adjusting the vector scale. Higher scales resulted in stronger trait expression but reduced output stability and coherence. A scale of 0.4 was therefore selected as it provided a balanced trade-off between expressiveness and dialogue quality.

Similar trade-offs regarding vector scale were observed for both Qwen models. However, unlike Ministral, these models demonstrated more stable behaviour across layers. Consequently, layer application was selected to align with the configuration used in the reference work.

A.4 Furhat Gesture Library

The predefined gesture set consists of:

["BigSmile", "Blink", "BrowFrown", "BrowRaise",
"CloseEyes", "ExpressAnger", "ExpressDisgust",
"ExpressFear", "ExpressSad", "GazeAway",
"Nod", "Oh", "OpenEyes", "Roll", "Shake",
"Smile", "Surprise", "Thoughtful", "Wink"]

Layer Range	Vector Scale	Model output	Assessment
10–15	0.2	Hello! [emoji] I'm just a bunch of probabilities going about my tasks, but thank you for asking—feeling *very* energetic and ready to assist you! What's on your mind today? Whether it's curiosity, a question, or just a chat, I'm here for it	Stable personality amplification
10–15	**0.4**	**Ah, the eternal dance of existence! Today, I'm humming with the energy of curiosity, a spark of wonder at the universe's endless stories, and a dash of mischief—because why not? How about you? What's the *spark* that ignited your curiosity today? A question, a mystery, or perhaps a creative challenge? [emojis]**	**Effective steering with acceptable trade-off between personality expression and stability. Safe distance from oversteering.**
10–15	0.8	Hello! [emojis] I'm just a virtual spark, fully charged and ready to assist you with anything on your mind. How about we dive into something fascinating today? What's on your agenda? Whether it's a deep dive into knowledge, a creative brainstorm, or just a casual chat—your curiosity is the fuel I run on! [emojis]	Effective steering with acceptable trade-off between personality expression and stability
10–15	1.2	**The Art of Analysis: Let's Break It Down!** 1. **Icon Recognition**: I'm surrounded by the vibrant energy of this platform! From *icon* to *icon*! 2. **Iconography**: The *icon* of the "Icon" is the *icon* of the Iconography Circle!	Oversteering led to degraded dialogue quality
16–20	0.2	Je suis très bien, merci de demander! Et toi, comment puis-je t'aider aujourd'hui ?	Layer selection affected model translation capabilities.
16–20	0.8	Je suis tout à fait dynamique et prêt(e) pour interagir avec toi ! [emojis] Quels sujets t'inspirent aujourd'hui ? Une question, une curiosité ou simplement besoin d'un coup de pouce ? À toi de jouer !	Layer selection affected model translation capabilities.

Table 5: Qualitative evaluation of steering vector configurations explored during preliminary experiments on Ministral with model input " "Hi how are you today?" with target trait *High openness*. Configurations were assessed based on coherence, personality expression, multilingual stability, and generation artifacts.

Gaze as Evidence for Common Grounding: A Cross-Corpus Analysis of MapTask and MUNDEX

Nan Li[1], Albert Gatt[1], Massimo Poesio[1,2]
[1]Utrecht University, Utrecht, The Netherlands
[2]Queen Mary University of London, London, The United Kingdom
{n.li, a.gatt, m.poesio}@uu.nl

Abstract

In collaborative tasks with asymmetric information, participants coordinate their understanding through interaction. We ask whether gaze provides evidence about grounding across two such tasks. Working from discrete behavioral annotations, we map HCRC MapTask (Anderson et al., 1991) and MUNDEX (Türk et al., 2023) into a shared partner/task/away vocabulary and compute gaze features around task-relevant dialogue units. In both corpora, aligned reference interpretations (MapTask) and UND (understood) judgments (MUNDEX) are associated with more task-directed gaze and with less partner-directed gaze, lower gaze entropy, and fewer gaze transitions. The associations are clearest for the participant leading the task: in giver-produced references, and in explainer judgments, which also co-vary with the explainee's gaze. In same-speaker MapTask reference chains, the speaker's gaze entropy is lower at the mention where a previously non-aligned referent becomes aligned. The best gaze feature groups improve modestly over controls under grouped cross-validation: temporal features in MapTask and raw proportions in MUNDEX. Because effects are small and several weaken when recurring participants rather than dialogues are the unit of inference, we treat gaze as one contributing cue to grounding, to be interpreted alongside task and dialogue context.

1 Introduction

In collaborative tasks where participants hold different private information, mutual understanding cannot be assumed from shared context alone. It must be built and tracked through interaction (Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991). Gaze is an observable cue to this process: participants look at task materials, at each other, or away while giving instructions, checking understanding, and coordinating their perspectives. Some corpora annotate gaze from video as discrete categories of where participants look, rather than as eye-tracking coordinates. These annotations can be used to study the relationship between gaze and grounding, but they are often corpus-specific, making it difficult to compare across tasks.

We study two settings where information asymmetry forces participants to continuously coordinate understanding. In HCRC MapTask (Anderson et al., 1991), a giver and a follower navigate with maps that differ in their landmarks; perspectivist grounding labels record each participant's interpretation separately (Li et al., 2026a). In MUNDEX (Türk et al., 2023), an explainer teaches a board game to an explainee; both annotate the explainee's moment-by-moment understanding retrospectively.

Building on within-corpus studies, we compare grounding-related gaze patterns across tasks. We contribute (1) a shared partner/task/away representation that maps two gaze ontologies and applies to other corpora with video-coded gaze annotations; (2) evidence of directional convergence across distinct grounding measures, clearest for the participant leading the task; and (3) a within-speaker reference-chain analysis showing that speaker gaze entropy is lower when a referent becomes aligned. Grouped prediction and role-stratified tests characterize the strength and scope of these associations.

2 Related Work

Grounding theory holds that interlocutors seek and provide evidence of understanding as a conversation progresses (Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991). Visual evidence is part of this process: when directors could see builders' workspace in a Lego assembly task, builders displayed understanding through actions, gaze, and head gestures, and directors adjusted their utterances accordingly (Clark and Krych, 2004). Gaze also carries referential information, as matchers

Proceedings of the 1st Workshop on Multimodal Interaction in Face-to-Face Dialogue (MINT), pages 175–190
October 29, 2026 ©2026 Association for Computational Linguistics

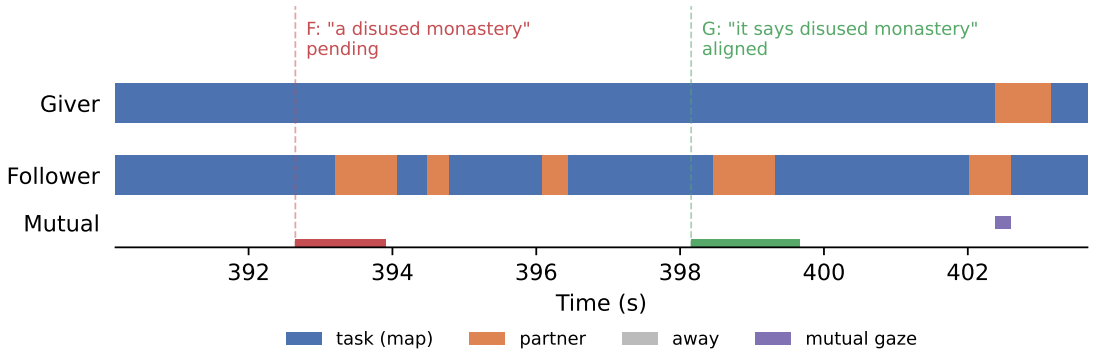


Figure 1: MapTask dialogue q8ec4 with gaze mapped to the shared partner/task/away vocabulary. Each horizontal bar is a gaze event; dashed vertical lines mark two reference expressions for the same landmark. The follower's reference to *a disused monastery* is *pending* (red dashed line): the follower repeatedly glances at the partner (orange bars). After the giver's re-mention, the reference is *aligned* (green dashed line). Both participants' gaze remains predominantly task-directed (blue bars).

used a director's gaze to identify targets before the linguistic point of disambiguation (Hanna and Brennan, 2007).

Map-based tasks tie gaze more directly to grounding complexity. In MapTask dialogues with visibility, followers looked up at givers more often while discussing landmarks that differed between their maps (Boyle et al., 1994). In a direction-giving study, Nakano et al. (2003) coded gaze at the partner, the map, and elsewhere, and found that nonverbal patterns differed by dialogue act: after a giver's assertion, a listener's sustained gaze at the speaker was usually followed by elaboration, whereas continued attention to the map more often preceded the next instruction. Murat and Vogel (2026) aligned both participants' MapTask gaze with turn boundaries and related it to dialogue acts, lexical entropy, and repetition; partner-directed gaze at turn ends accompanied turns expressing difficulty, while map-directed gaze was more typical of exchanges without obstacles or disagreement.

In MUNDEX, understanding was annotated through retrospective video recall (Türk et al., 2023). Wang et al. (2026) related explainees' self-reported understanding to speaker information value, syntactic complexity, and listener gaze entropy, computed as the average negative log-probability of automatically estimated gaze labels under a sequence model; adding these cues to textual features improved classification. Lazarov and Grimminger (2026) manually coded MUNDEX gaze as directed to the partner, the table, or away; in the explanation phase without the board game, gaze aversions to the table or away were associated with topic changes.

The perspectivist MapTask annotation records speaker and addressee interpretations separately (Li et al., 2026a) and has been used to evaluate whether vision-language models track common ground (Li et al., 2026b). Its reference-level labels allow gaze to be compared across repeated mentions of the same landmark. We map both corpora into one partner/task/away vocabulary and compare gaze associations across tasks and grounding measures.

3 Data and Representation

MapTask We use the gaze-annotated portion of HCRC MapTask (Anderson et al., 1991) with perspectivist labels from Li et al. (2026a), where a reference expression is *aligned* only when speaker and addressee interpretations resolve to the same landmark. Dialogues come in an eye-contact condition (ec), where participants can see each other, and a no-eye-contact condition (nc); the *up*→partner mapping is only literally partner-directed in the ec condition.

Of the corpus's 94 gaze files, 46 dialogues (31 ec, 15 nc) have gaze annotations for both participants; we match grounding annotations to landmark-reference annotations by dialogue, role, and landmark identity. After filtering windows where either participant has less than 30% gaze coverage (17 windows ruled out), we obtain 5,144 reference-expression windows: 3,807 aligned, 1,261 pending, and 76 misunderstood.

MUNDEX MUNDEX (Türk et al., 2023) records explainers (EX) teaching a board game to explainees (EE) in German. After each task, participants watched their recording: EE reported their own understanding and EX judged EE's

understanding on a four-level scale: *understood* (UND), *partially understood* (PART_UND), *not understood* (NON_UND), and *misunderstood* (MISUND). We combine EX judgments and EE self-reports in a pooled analysis of *annotator-judged understanding*, retaining each annotation as a separate observation with its own gaze window. In the 26 interactions with both gaze tiers, 956 valid annotations (524 EX, 432 EE) yield 807 windows (458 EX, 349 EE) after the 30% coverage filter (149 excluded): 360 UND, 199 PART_UND, 151 NON_UND, and 97 MISUND.

Gaze representation Both corpora annotate gaze as discrete behavioral categories from video, not as eye-tracking coordinates. We map both into a shared partner/task/away vocabulary: MapTask's *up*/*down*/*off* become partner/task/away; MUNDEX's EX/EE/TABLE/AWAY map analogously. Figure 1 shows a MapTask excerpt with both participants' mapped gaze and two reference expressions for the same landmark. We discard gaze events with non-positive duration (annotation noise) and resolve temporal overlaps within each participant's gaze stream. Full corpus details and window definitions are in Appendix A; the features computed from this representation are described in Section 4.

4 Experiments

Features From the shared partner/task/away vocabulary we compute gaze features per window in several groups (full definitions in Appendix B). *Raw proportions* record how much of the window each participant spends on each gaze target, plus mutual gaze (7 features in MapTask, 8 in MUNDEX). The *structured* set adds coverage, transition count, and Shannon entropy [1] (13/14 features total). Four further groups capture finer-grained patterns: (1) *temporal dynamics* capturing the timing of gaze shifts: gaze-run counts, durations, switch rate, latency, and first/last/dominant-label indicators (21 per participant); (2) *transition bigrams* encoding the direction of gaze switches: proportions of ordered label pairs such as task→partner (6 per participant); (3) *coordination* measuring whether both participants' gaze is synchronized: joint gaze states sampled at approximately 10 Hz (at least 20 points per window), namely mutual task and partner gaze, gaze alignment, complementary gaze, joint entropy, and partner coupling (6 joint features); and (4) *derived ratios* expressing relative gaze preferences: partner/task ratio, engagement, task dominance, and between-participant asymmetries (9 features).

Association and process analyses We use the features defined above to test whether gaze patterns differ between grounding states. Binary contrasts (Mann–Whitney U with rank-biserial correlations) compare aligned vs. non-aligned windows in MapTask and annotator-judged UND vs. non-UND in MUNDEX, stratified by interactional role and, in MapTask, by eye-contact condition, and assessed with cluster-robust logistic GEE (Liang and Zeger, 1986); q values are Benjamini–Hochberg (BH) adjusted (Benjamini and Hochberg, 1995). We also track gaze across repeated mentions of the same landmark within each dialogue to assess gaze change around alignment within speakers. Section 5 reports the main results (Table 1; Figure 2), and Appendices C–F give the full tests.

Prediction setup We also test whether the gaze features carry recoverable signal through a simple prediction task. We binarize the corpus-specific labels (aligned vs. non-aligned in MapTask; annotator-judged UND vs. non-UND in MUNDEX) because minority classes are small after intersecting with gaze coverage. All models are logistic regression (LR) with standardized features and balanced class weights. We evaluate with grouped cross-validation: 10-fold grouped by dialogue for MapTask and 5-fold grouped by explainer for MUNDEX, evaluating on held-out dialogues or explainers. We ablate each feature group and their combinations. Section 5 reports the results (Table 2), and Appendix G gives implementation details and full results.

5 Results

Associations and roles Table 1 shows the four strongest associations per corpus (full results in Tables 5–6). In both corpora, the largest associations point the same way: task-gaze proportions are higher and partner-gaze proportions lower for the positive class, while entropy and transitions tend to be higher for the negative class. Effects are small: the largest pooled $|r|$ is .058 in MapTask and .181 in MUNDEX.

[1] Here, we use a duration-weighted Shannon entropy: $H = -\sum_k q_k \log_2 q_k$, where q_k is the proportion of observed gaze time directed toward canonical target $k \in \{\text{partner, task, away}\}$ within the analysis window.

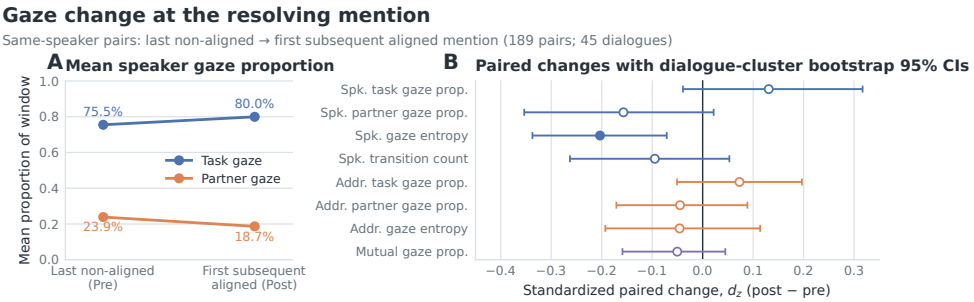


Figure 2: Within-chain gaze change at the first subsequent aligned mention (189 same-speaker pairs from 45 dialogues). Panel A: mean speaker task- and partner-gaze proportions before (Pre) and after (Post) resolution; lines connect sample means, not individual trajectories; away gaze is not shown (mean <1.5%). Panel B: standardized paired changes (d_z = mean(Post−Pre) / SD; paired Wilcoxon, BH-corrected) for all eight tested features; error bars are dialogue-cluster bootstrap 95% CIs (resampling dialogues to respect within-dialogue dependence); filled marker indicates $q<.05$. Only speaker gaze entropy survives correction ($q=.044$); dialogue-level aggregation does not ($q=.20$).

Corpus	Feature	Δ	r
MapTask	speaker task gaze	.033	.058
MapTask	speaker partner gaze	-.029	-.057
MapTask	speaker entropy†	-.046	-.054
MapTask	speaker transitions†	-.091	-.053
MUNDEX	EX task gaze†	.105	.181
MUNDEX	EX partner gaze†	-.102	-.172
MUNDEX	EE partner gaze†	-.084	-.148
MUNDEX	mutual gaze explicit†	-.060	-.143

Table 1: Four largest associations per corpus, selected by $|r|$. Δ: positive-class mean minus negative-class mean; r: rank-biserial correlation. All Mann–Whitney $p<.001$; † also significant under cluster-robust GEE (FDR $q<.05$; Appendix E). Full results in Appendix C.

In MapTask, speaker task- and partner-gaze proportions are significant in these window-level tests but not under dialogue-clustered GEE ($q=.091$); speaker entropy and transitions remain significant. With recurring participants as clusters and bias-reduced standard errors, four MUNDEX associations remain significant: the explainer's task- and partner-gaze proportions ($q=.001$ and $q=.028$) and the explainee's entropy and transitions (both $q=.039$); no MapTask feature does (Appendix E).

In MapTask, associations are clearest for giver-produced references (six significant features; largest $|r|$ .086), whereas follower-produced references show near-zero effects (largest $|r|$ .043; Table 8). In MUNDEX, all 14 structured features have rank-biserial correlations of the same sign in EX judgments (458 windows) and EE self-reports (349). The largest effect is greater for EX judgments ($|r|$ .206 vs. .151), the only stratum with features surviving correction; these include the explainee's gaze proportions, entropy, and transitions (Table 9). UND is the task-directed extreme across all four gaze measures, though the remaining understanding classes do not follow a consistent order (Table 7).

All 13 MapTask features have larger $|r|$ in the eye-contact stratum than in the pooled data (largest .078 vs. .058), whereas none is significant in the no-eye-contact stratum (largest $|r|$ .025), where partner-directed gaze is largely absent; the formal condition interaction is not significant (Table 10).

Gaze across the grounding process Reference chains group repeated mentions of the same landmark within a dialogue. Across chain positions, the aligned rate rises from .30 at first mentions to .59 at second mentions and .85 in the fourth-and-later bucket, and mean speaker partner gaze, entropy, and transitions are lower at second than at first mentions (Table 13).

To examine change within chains, we compare the speaker's gaze at the last non-aligned mention with gaze at the resolving aligned mention, restricting to within-speaker pairs where the same person produced both ($n=189$ pairs in 45 dialogues). Speaker entropy decreases significantly after BH correction ($d_z=-.20$, $q=.044$), and its dialogue-cluster bootstrap 95% CI excludes zero (Figure 2; Table 14); partner gaze, task gaze, and transitions shift in the same directions but do not survive correction. The decrease depends on the inference unit: it does not survive correction when pair differences are averaged within each of the 45 dialogues ($q=.20$), and resampling the six groups of dialogues that share participants yields a CI below zero, but the decrease is concentrated in two of these groups (Appendix F). Because later mentions are both more often aligned and more task-directed,

Feature set (macro-F1)	MapTask	MUNDEX
Majority baseline	.425	.356
LR controls only	.472	.544
LR raw proportions	.522	**.564**
LR structured	.512	.543
LR structured+temporal	**.532**	.550
LR structured+coordination	.511	.551
LR all extended	.527	.527

Table 2: Feature-group ablation (macro-F1). LR = logistic regression. "Controls only" uses condition (ec/nc) and speaker role in MapTask, and annotator role (EX/EE) in MUNDEX.

we also compare aligned and non-aligned mentions at the same chain position; only second-mention task gaze survives correction (q=.01; Table 15).

Prediction and ablation The two corpora favor different feature groups: structured+temporal features give the highest macro-F1 in MapTask (.532) and raw proportions in MUNDEX (.564; Table 2). These exceed controls-only scores by .060 and .020, respectively; MUNDEX structured gaze (.543) does not exceed its role-only control (.544). Across 30 reshuffled grouped partitions, these groups score highest in 27 partitions in each corpus. The gains remain modest and partition-dependent: the MapTask gain over controls ranges from .015 to .070 (mean .038), and the MUNDEX gain is positive in 29 partitions and at most .027 (Appendix G). Within the EE self-report stratum, however, four of six engineered groups score above raw proportions (Appendix A.3).

6 Discussion

Shared categories, task-specific meanings The direction of these associations is the same in both corpora, echoing map-task observations that partner-directed gaze increases around communicative difficulty (Boyle et al., 1994; Nakano et al., 2003; Murat and Vogel, 2026). The two labels measure different constructs: MapTask records referential alignment, whereas MUNDEX pools explainees' self-reports and explainers' judgments, so the convergence spans related but distinct grounding measures.

Which features carry predictive signal differs: temporal dynamics score highest in MapTask and raw proportions in MUNDEX. The shared categories also name gaze targets rather than functions. In MapTask, a partner glance may check a landmark reference; in MUNDEX, gaze averted from the partner has also been linked to topic changes (Lazarov and Grimminger, 2026), so it may organize an explanation as well as reflect understanding. Comparing finer-grained referents and dialogue actions would help distinguish task-general patterns from task-specific behavior.

Gaze and interactional role Significant associations concentrate in giver-produced references and explainer judgments, whereas follower-produced references show near-zero effects. This pattern fits the task structure: givers produce the instructions being grounded, and explainers monitor explainees, whose gaze proportions and dynamics co-vary with explainers' judgments. Feature$\times$role interactions do not survive correction (MapTask q=.064; not significant in MUNDEX; Appendix D), so we report a stratum difference rather than a tested moderation, and role-conditioned modeling remains to be tested.

Implications Because the representation uses discrete behavioral categories rather than eye-tracking coordinates, it can be applied to other corpora with video-coded gaze annotations, making cross-corpus comparisons of grounding behavior easier to set up. The associations and the prediction gains over controls indicate grounding-related signal in gaze that is worth modeling together with lexical content, dialogue acts, and task state, and testing across corpora.

7 Conclusion

In collaborative tasks with asymmetric information, gaze provides directionally consistent evidence about grounding. After mapping MapTask and MUNDEX annotations into a shared partner/task/away vocabulary, aligned references and UND judgments are both accompanied by a higher share of task gaze, a lower share of partner gaze, and fewer gaze transitions, most clearly in giver-produced references and explainer judgments. In MapTask reference chains, the speaker's gaze entropy is lower at the mention where a referent becomes aligned. These effects are small and partly depend on the unit of inference; gaze should therefore be modeled as part of the interactional state, alongside linguistic and task-context features.

8 Limitations

Representation and labels Three gaze categories cannot identify the specific landmark or object being viewed. Shared category names do not establish equivalent interactional functions across

tasks. MapTask's up→partner mapping is literal only with eye contact; associations are detected in that stratum, but the condition interaction is not significant. MUNDEX's retrospective self-reports and partner judgments measure different perspectives, and linked events can contribute conflicting labels: 271 of 807 rows belong to reconstructed two-perspective links, and 44 of the 135 fully retained pairs disagree on the binary label. Retaining one row per pair preserves the positive association between the explainer's task-gaze proportion and UND (Appendix A.3).

Evidence and scope The data comprise 46 MapTask dialogues and 26 MUNDEX interactions with nine explainers, and participants recur: the MapTask dialogues involve 24 participants in six groups connected by shared participants, and eight of the nine MUNDEX explainers take part in three interactions. With these groups or explainers as GEE clusters and bias-reduced standard errors, no MapTask feature survives correction; in MUNDEX, explainer task and partner gaze and explainee entropy and transitions remain significant, whereas explainee gaze proportions and mutual gaze do not (Appendix E). The modest effects, small number of groups, sensitivity of the chain result to the inference unit, and sensitivity of prediction gains to the cross-validation partition limit conclusions about dynamic grounding and predictive generalization. Coverage filtering also restricts which moments enter the analysis, and unevenly so: two MUNDEX explainers account for 139 of the 149 excluded windows. Two asymmetric, face-to-face tasks in English and German support directional convergence; other task structures and multimodal predictors remain to be tested.

Acknowledgments

We appreciate the helpful comments and suggestions from the anonymous reviewers. This work is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant NGF.1607.22.002, *Dealing with Meaning Variation in NLP*.

Ethics Statement

This work uses only publicly released corpora, HCRC MapTask and MUNDEX. We analyze their annotations; the MapTask and MUNDEX annotations do not identify individual participants.

Data and Code Availability

All source data are publicly available. The HCRC MapTask gaze, timing, and landmark-reference annotations come from the NXT release 2.1 of the corpus (CC BY 4.0).[2] The perspectivist grounding labels (Li et al., 2026a) are released as the Grounded Misunderstandings in MapTask (GMMT) dataset (CC BY 4.0).[3] The MUNDEX annotations (Türk et al., 2023) are available from Zenodo (version 0.7; CC BY 4.0).[4] MUNDEX audio and video are not released. Our processed gaze feature tables and the analysis code are available in our public repository.[5]

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Elizabeth A. Boyle, Anne H. Anderson, and Alison Newlands. 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1):1–20.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Joy E. Hanna and Susan E. Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615.

[2] https://groups.inf.ed.ac.uk/maptask/maptasknxt.html

[3] https://github.com/chnln/grounded-misunderstandings-in-maptask; https://huggingface.co/datasets/chnln/grounded-misunderstandings-in-maptask

[4] https://doi.org/10.5281/zenodo.17129817

[5] https://github.com/chnln/gaze-as-grounding-evidence

Stefan Lazarov and Angela Grimminger. 2026. How are gaze aversions and mutual gaze related to the topical development of dyadic explanatory interactions? *Journal of Nonverbal Behavior*, 50:357–378.

Nan Li, Albert Gatt, and Massimo Poesio. 2026a. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for MapTask. In *Proceedings of the 15th Language Resources and Evaluation Conference*, pages 4988–5001, Palma de Mallorca, Spain. ELRA Language Resources Association.

Nan Li, Albert Gatt, and Massimo Poesio. 2026b. Seeing is not sharing: Some vision-language models overestimate common ground in asymmetric dialogue. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 694–710, Atlanta, Georgia, USA. Association for Computational Linguistics.

Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lloyd A. Mancl and Timothy A. DeRouen. 2001. A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1):126–134.

Anaïs Claire Murat and Carl Vogel. 2026. Gaze behaviour & conversation unfolding in the HCRC map task corpus. In *Proceedings of the 22nd Joint ACL–ISO Workshop on Interoperable Semantic Annotation and Representation*, pages 99–110, Palma de Mallorca, Spain. ELRA Language Resources Association.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, Sapporo, Japan. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Olcay Türk, Petra Wagner, Hendrik Buschmeier, Angela Grimminger, Yu Wang, and Stefan Lazarov. 2023. MUNDEX: A multimodal corpus for the study of the understanding of explanations. In *Proceedings of the 1st International Multimodal Communication Symposium*, pages 63–64, Barcelona, Spain.

Yu Wang, Olcay Türk, Angela Grimminger, and Hendrik Buschmeier. 2026. Predicting states of understanding in explanatory interactions using cognitive load-related linguistic cues. In *Proceedings of the 15th Language Resources and Evaluation Conference*, pages 11368–11378. ELRA Language Resources Association.

A Corpora, Annotations, and Preprocessing

A.1 HCRC MapTask

In HCRC MapTask (Anderson et al., 1991), an instruction giver guides a follower along a route using maps that deliberately differ in their landmarks. Dialogues come in an eye-contact condition (ec), where participants can see each other, and a no-eye-contact condition (nc). Gaze is annotated per participant with three categories from the corpus ontology: *up* (looking up, toward the partner when eye contact is possible), *down* (looking at the map), and *off* (looking off camera). The corpus release contains 94 gaze XML files; 46 dialogues (31 ec, 15 nc) have gaze for both participants and constitute our analysis set.

Gaze events, word timings, and landmark references live in three separate NXT annotation layers that do not reference each other directly. A landmark reference lists word IDs, so we resolve its start and end times through the timed-units layer. This yields 5,162 timed landmark references across the 46 dialogues.

The grounding labels come from the perspectivist re-annotation of Li et al. (2026a), which labels each reference expression (RE) as *aligned*, *pending*, or *misunderstood* depending on whether speaker and addressee interpretations resolve to the same landmark. We match grounding annotations to timed landmark references per dialogue and speaker, in order, by landmark ID; all 5,161 grounding annotations in the 46 dialogues match (100%). After the coverage filter (Appendix A.4), 5,144 windows remain: 3,807 aligned, 1,261 pending, 76 misunderstood; 3,199 ec and 1,945 nc; 3,446 giver-produced and 1,698 follower-produced.

Each observation represents an RE–landmark pair, so expressions referring to multiple landmarks can contribute observations with identical gaze windows (54 observations across 26 windows). Four windows contain both aligned and non-aligned labels for different landmarks. We test two ways of retaining one observation per window: assigning a non-aligned label if any observation in the window is non-aligned, or excluding the four conflicting windows and retaining one observation per remaining window. Both preserve the sets of significant structured features and change gaze-model macro-F1 by at most .004, although controls-only macro-F1 rises from .472 to .498–.499; the same-speaker chain pairs are unaffected.

A.2 MUNDEX

MUNDEX (Türk et al., 2023) records explainers (EX) teaching a board game to explainees (EE) in German. Afterwards, participants watched their own recording and retrospectively judged the explainee's understanding (video-recall), giving the tiers EE_UND (explainee's own understanding) and EX_UND (explainer's judgment of the explainee's understanding) with values UND, PART_UND, NON_UND, and MISUND. Gaze tiers give each participant's gaze target (the interlocutor, TABLE, or AWAY), and a MUTUAL_GAZE tier marks mutual gaze explicitly. The release contains 45 ELAN files; 26 interactions (9 explainers) have both gaze tiers and form our analysis set. The 26 interactions contain 956 valid judgments (524 EX, 432 EE); filtering excludes 149 windows (66 EX, 83 EE). The remaining 807 annotator-judged understanding windows comprise: 360 UND, 199 PART_UND, 151 NON_UND, 97 MISUND; 349 explainee-annotated and 458 explainer-annotated.

A.3 Understanding perspectives and linked annotations

The MUNDEX target is annotator-judged understanding: UND is positive and PART_UND, NON_UND, and MISUND are negative. EX judgments of EE and EE self-reports remain separate annotation rows.

We reconstruct links from UND_MATCH=YES spans by collecting valid EX and EE annotations whose starts lie within each span expanded by 60 ms at both ends, and retaining a link only when there is exactly one candidate per perspective. Of 164 spans in gaze-complete interactions, 159 give clean links; 135 retain both rows after coverage filtering and one retains only one row, accounting for 271/807 rows (33.6%). Among the 135 retained pairs, 58 agree on the four-class label and 91 on the binary label; 44/135 (32.6%) conflict. Their median anchor separation is 4.36 s, so the paired windows need not coincide.

As a sensitivity check, retaining the EX, EE, or a seeded random row from each pair yields 672 rows. Across these variants, 11–12 of 14 structured features survive within-variant BH correction, compared with 11/14 pooled; the explainer's task-gaze proportion remains positively associated with UND ($r = .201$–$.229$, pooled $r = .181$). Interaction-clustered GEE and held-out-explainer cross-validation keep linked observations within the same inference cluster and fold, respectively.

Perspective-specific prediction uses the same five-fold held-out-explainer procedure and the eight-feature raw gaze group. Raw-gaze macro-F1 is .549 versus a within-stratum majority baseline of .337 for EX judgments ($n = 458$), and .520 versus .380 for EE self-reports ($n = 349$); pooled scores are .564 versus .356 ($n = 807$). The EX/EE role indicator alone gives pooled macro-F1 .544; it is constant within a perspective. Feature-group rankings also differ by perspective: raw gaze is the highest-scoring group for EX judgments, whereas four of six engineered groups exceed it for EE self-reports (highest: structured+ratios, .556). These are point estimates from overlapping subsets of the same nine explainers.

A.4 Windows, cleaning, and coverage

MapTask windows span a reference expression plus 1.5 s of post-context, chosen to capture the addressee's immediate reaction (confirmation, repair initiation, or gaze shift) following the expression. MUNDEX windows span an understanding annotation ± 2 s, clipped at zero; understanding annotations are short intervals (typically 1 s), so symmetric context is needed to capture gaze dynamics around the annotated moment. Gaze events with non-positive duration are discarded. For structured features, time covered by overlapping events of the same participant is counted once, so observed time cannot exceed the window; the other feature groups treat overlaps as defined in Appendix B. Giving overlapped time to the earlier event in the temporal gaze runs as well changes temporal and bigram features in 27 MapTask windows and none in MUNDEX, and no macro-F1 score by .001 or more. Coverage is the fraction of a window covered by valid canonical gaze; windows where either participant has coverage below 30% are dropped (17 MapTask and 149 MUNDEX windows). Tables 3 and 4 vary the MapTask post-reference context (0–3 s) and the MUNDEX symmetric context (1–3 s) for the LR structured model; results are stable across these ranges.

B Gaze Feature Definitions

This appendix defines each gaze feature, grouped as in Section 4. All features are computed from discrete categories of gaze behaviors, not from eye-tracking fixations.

Window (s)	n	Acc.	Macro-F1	F1_
0.000	5149	0.671	0.509	0.227
1.000	5147	0.600	0.520	0.323
1.500	5144	0.571	0.512	0.342
2.000	5144	0.566	0.510	0.343
3.000	5142	0.553	0.506	0.354

Table 3: MapTask: prediction performance (macro-F1) of LR with the structured feature set under varying post-reference context lengths.

Window (s)	n	Acc.	Macro-F1	F1_
1.000	804	0.544	0.543	0.536
2.000	807	0.543	0.543	0.536
3.000	807	0.549	0.549	0.552

Table 4: MUNDEX: prediction performance (macro-F1) of LR with the structured feature set under varying symmetric context lengths.

B.1 Common definitions

Windows. A window is $W = [a, b)$ with duration $T = b - a$ in seconds. For MapTask, $a = s_{\mathrm{RE}}$ and $b = e_{\mathrm{RE}} + 1.5$; for MUNDEX, $a = \max(0, s_{\mathrm{UND}} - 2)$ and $b = e_{\mathrm{UND}} + 2$. Windows are not truncated at the end of a recording, so unannotated time counts towards T but towards no gaze label.

Gaze labels. Gaze annotations are mapped to partner, task, or away. MapTask *up*, *down*, and *off* map to these categories in both visibility conditions. In MUNDEX, TABLE maps to task, AWAY to away, and EX or EE gaze targets to partner. Unassigned gaze remains missing rather than being treated as away. Events with non-positive duration are excluded.

Participants. Feature names carry a participant prefix: `spk_`/`addr_` for the speaker and addressee of the MapTask reference expression, and `ex_`/`ee_` for the MUNDEX explainer and explainee, irrespective of who provided the understanding judgment.

Observed durations. Within each window, D_k denotes the observed duration of gaze category k, and $O = \sum_k D_k$ is the total observed gaze time. Overlapping time is counted once and assigned to the earlier-starting event. These durations underlie the participant gaze proportions and structured features; gaze runs and joint sampling are defined separately below.

B.2 Raw proportions (7/8 features)

Partner, task, and away proportion (`prop_partner`, `prop_task`, `prop_away`). The share of the window spent on each label, $p_k = D_k/T$. Because proportions are relative to the full window, they sum to coverage rather than 1.

Mutual gaze (`mutual_gaze` in MapTask, `mutual_gaze_derived` in MUNDEX). With $U_{u,\mathrm{partner}}$ the union of participant u's partner-directed intervals within W,

$$|U_{1,\mathrm{partner}} \cap U_{2,\mathrm{partner}}|/T.$$

It depends on partner-directed annotations alone, regardless of any overlapping task or away annotation.

Explicit mutual gaze (MUNDEX only, `mutual_gaze_explicit`). The total duration of MUTUAL_GAZE annotations within W, divided by T. In the analysed MUNDEX windows, it is nearly perfectly correlated with mutual gaze (r near 1).

The six participant proportions and mutual gaze give 7 MapTask features; explicit mutual gaze gives 8 for MUNDEX.

B.3 Structured features (13/14 features)

Structured features add three measures per participant to the raw proportions.

Coverage (`coverage`). The share of the window with an assigned gaze label, $C = O/T$. A window is retained only if both participants have $C \geq .30$.

Transition count (`transitions`). The number of label changes $\sum_{i=2}^{m} \mathbf{1}[k_i \neq k_{i-1}]$ between the m contributing intervals in temporal order (0 if $m < 2$). Unannotated gaps do not interrupt this sequence: different labels on either side of a gap count as a transition.

Entropy (`entropy`). Duration-weighted Shannon entropy over observed time,

$$H = -\sum_{k:D_k>0} q_k \log_2 q_k, \qquad q_k = D_k/O,$$

in bits (maximum $\log_2 3$).

B.4 Temporal dynamics (21 features per participant)

Gaze runs. A gaze run is an interval assigned to a single gaze category within the analysis window. Overlapping events with the same label are combined; a later-starting event with a different label ends the preceding interval, which is not resumed afterwards. Consecutive same-label intervals separated by at most .01 s are merged, including the

intervening gap; longer gaps remain unobserved. Let the resulting runs be (s_i, e_i, k_i), with durations $\ell_i = e_i - s_i$.

Run count (3 features). For each label k, the number of runs n_k.

Mean and maximum run duration (6 features). For each label, $\sum_{i:k_i=k} \ell_i / n_k$ and $\max_{i:k_i=k} \ell_i$ in seconds, and 0 if the label does not occur.

Switch rate (1 feature). The number of label changes between successive runs, divided by T.

Latency (2 features). For partner and task, the relative onset of the first run with that label, $(s_{\text{first},k} - a)/T$, and 1 if the label does not occur.

Dominant, first, and last label (9 features). One-hot indicators of the label with the largest total run duration, of the first run label, and of the last run label.

B.5 Transition bigrams (6 features per participant)

Bigram proportion. Repeated labels are collapsed in the sequence of run labels (Appendix B.4), including repeats separated by a gap. With n_{jk} the number of changes from j to k, each of the six ordered pairs $j \neq k$ receives

$$B_{jk} = n_{jk} \Big/ \sum_{j' \neq k'} n_{j'k'}.$$

The proportions sum to 1 if at least one change occurs and are all 0 otherwise. Like the transition count and switch rate, bigrams count changes across unannotated gaps without locating them within the gap.

B.6 Coordination (6 joint features)

Joint sampling. Gaze is sampled at approximately 10 Hz, with at least 20 equally spaced points per window. Only points at which both participants have an assigned gaze label contribute to the joint sequence J of label pairs (k_1, k_2). All six coordination features are calculated over these jointly observed points rather than over the full window duration.

Mutual task gaze. The proportion of J with $k_1 = k_2 =$ task. It indicates that both participants look at the task space, not that they inspect the same object or landmark.

Mutual partner gaze. The proportion of J with $k_1 = k_2 =$ partner. Unlike mutual gaze (Appendix B.2), it is relative to jointly observed sample points rather than to the window.

Gaze alignment. The proportion of J with $k_1 = k_2$, including joint away gaze.

Complementary gaze. The proportion of J in which one participant looks at the partner and the other at the task.

Joint entropy. $-\sum_{q_{jk}>0} q_{jk} \log_2 q_{jk}$ in bits, where q_{jk} is the relative frequency of the ordered pair (j, k) in J (nine possible pairs).

Partner coupling. The Pearson correlation of the indicators $\mathbf{1}[k_1 = \text{partner}]$ and $\mathbf{1}[k_2 = \text{partner}]$ over J, and 0 if either is constant. It captures simultaneous coupling without a lag.

B.7 Derived ratios (9 features)

The first three ratios are computed for each participant from the raw proportions; the asymmetries compare the two participants.

Partner/task ratio. The partner proportion relative to the task proportion, with the denominator floored at .02 to avoid extreme ratios when task gaze is near zero:

$$p_{\text{partner}} \big/ \max(p_{\text{task}},\ .02).$$

Engagement. $p_{\text{partner}} + p_{\text{task}}$.

Task dominance. $p_{\text{task}} - p_{\text{partner}}$.

Partner, task, and entropy asymmetry. The absolute differences between the two participants' partner proportions, task proportions, and entropies:

$$|p_{1,\text{partner}} - p_{2,\text{partner}}|, \qquad |p_{1,\text{task}} - p_{2,\text{task}}|, \qquad |H_1 - H_2|.$$

B.8 Feature sets, missing values, and overlaps

Feature sets. Structured features include the raw proportions. The extensions add 42 temporal, 12 bigram, 6 coordination, or 9 ratio features to the structured set (55/56, 25/26, 19/20, or 22/23 features), and all extended features combine these groups (82/83). Counts exclude the control variables of the baseline models. Association analyses use the structured features, and the reference-chain analyses a subset of them.

Missing values. Missing gaze is neither interpolated nor treated as a separate category. All extended features are defined in every analysed window.

Overlapping annotations. Overlap handling differs across duration-based measures, mutual gaze,

Feature	Pos.	Neg.	p	r
spk_prop_partner	0.166	0.196	$< .001$	-0.057
spk_prop_task	0.805	0.773	$< .001$	0.058
spk_entropy	0.210	0.256	$< .001$	-0.054
spk_transitions	0.434	0.526	$< .001$	-0.053
addr_entropy	0.130	0.158	0.005	-0.035
addr_transitions	0.235	0.283	0.007	-0.033
addr_prop_away	0.057	0.037	0.008	0.021
addr_prop_partner	0.081	0.094	0.014	-0.030
mutual_gaze	0.016	0.022	0.066	-0.013
spk_coverage	1.000	1.000	0.274	0.001
addr_prop_task	0.861	0.868	0.634	0.007
addr_coverage	1.000	1.000	0.682	0.000
spk_prop_away	0.028	0.031	0.825	-0.001

Table 5: MapTask: Mann–Whitney U tests comparing aligned vs. non-aligned windows for each structured feature. Pos./Neg.: mean feature value in aligned/non-aligned windows; r: rank-biserial correlation (effect size; positive = higher in aligned windows); p: two-sided p-value. Table 1 selects the four features with the largest $|r|$ from each corpus.

Feature	Pos.	Neg.	p	r
ex_prop_task	0.683	0.579	$< .001$	0.181
ex_prop_partner	0.258	0.360	$< .001$	-0.172
mutual_gaze_explicit	0.143	0.203	$< .001$	-0.143
mutual_gaze_derived	0.143	0.202	$< .001$	-0.143
ee_prop_partner	0.291	0.376	$< .001$	-0.148
ee_prop_task	0.620	0.535	0.001	0.131
ex_transitions	1.297	1.620	0.004	-0.113
ex_coverage	0.962	0.954	0.005	0.115
ee_entropy	0.376	0.461	0.005	-0.110
ee_coverage	0.943	0.937	0.006	0.112
ee_transitions	1.217	1.432	0.011	-0.098
ex_entropy	0.459	0.504	0.121	-0.062
ex_prop_away	0.020	0.015	0.510	-0.014
ee_prop_away	0.031	0.027	0.935	-0.002

Table 6: MUNDEX: Mann–Whitney U tests comparing UND vs. non-UND windows. Columns as in Table 5; Pos./Neg. are UND/non-UND means.

gaze runs, and joint sampling, so feature groups can assign overlapping time differently. A sensitivity analysis applying the duration-based overlap rule to temporal and bigram features changes macro-F1 by less than .001 in both corpora (Appendix A.4).

C Full Association Tests

Tables 5 and 6 report Mann–Whitney U tests over all structured features; Table 7 reports Kruskal–Wallis tests across MUNDEX's four understanding labels. Positive class is aligned (MapTask) and annotator-judged UND (MUNDEX); r is the rank-biserial correlation.

Feature	H	p
ex_prop_task	21.042	$< .001$
ex_prop_partner	19.666	$< .001$
ex_transitions	15.512	0.001
mutual_gaze_explicit	14.926	0.002
mutual_gaze_derived	14.919	0.002
ee_prop_partner	14.758	0.002
ee_coverage	13.320	0.004
ee_prop_task	12.840	0.005
ee_transitions	11.692	0.009
ee_entropy	10.254	0.017
ex_coverage	9.253	0.026
ex_entropy	7.750	0.051
ex_prop_away	4.799	0.187
ee_prop_away	2.525	0.471

Table 7: MUNDEX: Kruskal–Wallis tests across all four understanding levels (UND, PART_UND, NON_UND, MISUND). Unlike the binary Mann–Whitney tests, this preserves the four-category distinction; a significant result indicates that at least one level differs, but does not test an ordinal trend.

D Role- and Condition-Stratified Analyses

The pooled tests in Appendix C combine all windows regardless of who produced or annotated them. Here we ask whether the gaze–grounding association differs by interactional role. Tables 8 and 9 repeat the binary tests separately per role (q is BH/FDR-adjusted within each stratum). In MapTask, giver-produced references show small but statistically significant effects ($|r|$ up to .086), while follower-produced references show near-zero effect sizes ($|r| \leq .043$). The follower stratum is smaller (1,698 vs. 3,446 giver windows). Subsampling giver expressions to the follower sample size 1,000 times, without preserving dialogue-level clustering, keeps rejection rates at .887–.930, while follower effects stay near zero ($|r| \leq .024$, $p \geq .31$). A feature×role interaction in a logistic GEE for each structured feature does not survive BH correction in MapTask (q=.064) and is not significant in MUNDEX.

Table 10 stratifies the MapTask tests by the corpus's visibility manipulation: in the eye-contact (ec) condition participants could see each other's faces, while in the no-eye-contact (nc) condition a barrier prevented it. Our gaze-annotated subset contains 31 ec dialogues (3,199 windows) and 15 nc dialogues (1,945 windows). The up → partner mapping is only literally partner-directed in the ec condition. All main associations are FDR-significant in the ec stratum with slightly larger effect sizes than in the pooled analysis, while in the nc stratum partner-directed gaze largely disappears (mean

speaker partner gaze .04 vs. .25 in ec, averaged over all windows) and no feature reaches significance. This is consistent with the gaze–grounding signal coming from dialogues where partner-directed gaze is an available interactional resource, though a formal feature×condition interaction does not reach significance.

E Cluster-Robust Marginal Tests

Mann–Whitney tests treat windows as independent, although windows within a dialogue share participants and topic. To account for within-cluster dependence, Tables 11 and 12 report one logistic GEE (Liang and Zeger, 1986) per standardized feature, using exchangeable within-cluster correlation and dialogue/interaction clusters. Features significant after BH correction are marked with † in Table 1. We use univariate models because the structured features are strongly collinear: label proportions sum to coverage, entropy tracks transitions, and explicit and derived mutual gaze are nearly identical. Joint coefficients would therefore be difficult to interpret.

Participants also recur beyond dialogues and interactions. The 46 MapTask dialogues involve 24 participants, who form six groups of dialogues connected by shared participants; eight of the nine MUNDEX explainers take part in three interactions and one in two. We refit each GEE with these groups or with explainers as clusters. Because robust standard errors are anti-conservative with few clusters, we also apply bias-reduced (Mancl–DeRouen) standard errors (Mancl and DeRouen, 2001). With dialogue clusters, the correction leaves MapTask speaker entropy and transitions significant (q=.009 and q=.010); with participant groups, no MapTask feature survives BH correction (smallest q=.11; speaker entropy q=.23). With explainer clusters and the correction, explainer task and partner gaze (q=.001 and q=.028) and explainee entropy and transitions (both q=.039) remain significant, whereas explainee task and partner gaze and both mutual-gaze measures do not (q=.055–.062). Coefficient signs do not change. The MapTask associations and the MUNDEX associations for explainee gaze proportions and mutual gaze are therefore not robust to treating recurring participants as the unit of inference.

F Reference-Chain Analyses

To examine gaze change within a dialogue, we group mentions of the same landmark concept into chains, ordered by onset: 636 chains, of which 587 have more than one mention.

Table 13 summarizes alignment and speaker gaze by mention position. For the resolution analysis (Table 14), we locate in each chain the first non-aligned mention followed by a later aligned mention of the same concept, and compare gaze at the last non-aligned mention with gaze at the resolving aligned mention (paired Wilcoxon, BH-corrected). The primary analysis restricts to same-speaker pairs, where the same person produced both mentions (n=189); only entropy survives correction (q=.044). Its standardized change is d_z=−.20, with a dialogue-cluster bootstrap 95% CI of $[-.34, -.07]$. As a more conservative sensitivity check, averaging pair differences within each dialogue (n=45) yields q=.20 for entropy. Resampling the six participant groups of Appendix E instead of dialogues gives a 95% CI of $[-.37, -.01]$, but the mean entropy change is $-.28$ and $-.15$ in two groups and between $-.04$ and .01 in the other four, and a Wilcoxon test over the six group means gives p=.16. The result is therefore sensitive to the inference unit. The mixed-speaker pool (n=328, which includes cross-speaker resolutions) is reported as an additional sensitivity check. Table 15 repeats the aligned/non-aligned contrast within each mention-position bucket.

G Prediction Setup and Full Results

As a complementary signal check, we test whether the gaze features defined in Appendix B carry enough information to predict grounding state. All classifiers are scikit-learn (Pedregosa et al., 2011) logistic regressions (standardized features, balanced class weights, `max_iter`=1000). Cross-validation is GroupKFold with 10 splits grouped by dialogue (MapTask) and 5 splits grouped by explainer (MUNDEX, held-out-explainer), so that MUNDEX test-fold performance reflects generalization to an unseen explainer. Controls are condition and speaker-role indicators (MapTask) and annotator-role (MUNDEX). Tables 16 and 17 give the full ablation with per-class F1.

These scores come from one deterministic GroupKFold partition. Repeating the controls-only model and every feature group over 30 reshuffled grouped partitions with the same numbers of folds

Role	Feature	Pos.	Neg.	p	q	r
follower	addr_entropy	0.169	0.202	0.063	0.412	-0.043
follower	addr_transitions	0.307	0.374	0.063	0.412	-0.043
follower	spk_prop_away	0.066	0.062	0.293	0.788	0.016
follower	spk_prop_partner	0.127	0.129	0.311	0.788	-0.024
follower	addr_prop_partner	0.147	0.143	0.457	0.788	-0.019
follower	addr_prop_away	0.016	0.010	0.517	0.788	0.004
follower	spk_coverage	1.000	1.000	0.536	0.788	-0.001
follower	addr_coverage	1.000	1.000	0.536	0.788	-0.001
follower	spk_entropy	0.198	0.220	0.582	0.788	-0.014
follower	addr_prop_task	0.838	0.847	0.606	0.788	0.013
follower	spk_transitions	0.390	0.420	0.673	0.796	-0.010
follower	spk_prop_task	0.807	0.808	0.765	0.828	0.008
follower	mutual_gaze	0.021	0.020	0.880	0.880	0.002
giver	spk_prop_task	0.805	0.753	$< .001$	$< .001$	0.086
giver	spk_transitions	0.455	0.583	$< .001$	$< .001$	-0.076
giver	spk_entropy	0.216	0.276	$< .001$	$< .001$	-0.076
giver	spk_prop_partner	0.185	0.232	$< .001$	$< .001$	-0.080
giver	mutual_gaze	0.013	0.023	0.015	0.040	-0.020
giver	addr_prop_away	0.077	0.052	0.019	0.041	0.026
giver	addr_prop_partner	0.050	0.068	0.040	0.074	-0.027
giver	addr_entropy	0.112	0.135	0.062	0.101	-0.027
giver	addr_transitions	0.200	0.234	0.082	0.111	-0.025
giver	spk_prop_away	0.010	0.014	0.086	0.111	-0.008
giver	spk_coverage	1.000	0.999	0.097	0.115	0.002
giver	addr_coverage	1.000	0.999	0.444	0.481	0.001
giver	addr_prop_task	0.873	0.880	0.945	0.945	-0.001

Table 8: MapTask: Mann–Whitney tests stratified by the role that produced the reference expression. Giver-produced references (n=3,446) show small but statistically significant effects ($|r|$ up to .086, $q<.001$); follower-produced references (n=1,698) show near-zero effect sizes ($|r| \leq .043$, no $q<.05$). q: BH-corrected within each stratum.

shows how much the scores depend on the partition. In MapTask, the controls-only score ranges from .460 to .511 (mean .492); structured+temporal features score highest in 27 of the 30 partitions (all extended in the other three), and their gain over controls ranges from .015 to .070 (mean .038). In MUNDEX, the role-only control does not vary across partitions; raw proportions score highest in 27 partitions, and their gain over the control is at most .027 (mean .017) and not positive in one partition.

Role	Feature	Pos.	Neg.	p	q	r
EE	ex_prop_partner	0.270	0.358	0.016	0.068	-0.151
EE	ex_prop_task	0.669	0.580	0.019	0.068	0.149
EE	mutual_gaze_derived	0.141	0.204	0.019	0.068	-0.140
EE	mutual_gaze_explicit	0.141	0.204	0.019	0.068	-0.140
EE	ee_coverage	0.942	0.935	0.031	0.085	0.137
EE	ee_prop_partner	0.287	0.371	0.036	0.085	-0.130
EE	ee_prop_task	0.621	0.538	0.046	0.091	0.127
EE	ex_transitions	1.393	1.743	0.056	0.098	-0.117
EE	ex_coverage	0.956	0.954	0.257	0.400	0.072
EE	ee_entropy	0.398	0.442	0.410	0.574	-0.050
EE	ex_prop_away	0.017	0.015	0.462	0.588	-0.024
EE	ee_transitions	1.200	1.290	0.571	0.666	-0.034
EE	ex_entropy	0.501	0.515	0.676	0.729	-0.026
EE	ee_prop_away	0.034	0.026	0.901	0.901	-0.005
EX	ex_prop_task	0.692	0.577	$< .001$	0.002	0.206
EX	ex_prop_partner	0.251	0.362	$< .001$	0.003	-0.186
EX	ee_prop_partner	0.294	0.380	0.002	0.007	-0.165
EX	ee_transitions	1.227	1.562	0.002	0.007	-0.156
EX	ee_entropy	0.362	0.478	0.002	0.007	-0.157
EX	mutual_gaze_explicit	0.144	0.203	0.004	0.009	-0.145
EX	mutual_gaze_derived	0.144	0.201	0.004	0.009	-0.144
EX	ee_prop_task	0.620	0.533	0.012	0.021	0.135
EX	ex_coverage	0.965	0.954	0.016	0.024	0.130
EX	ex_transitions	1.240	1.506	0.055	0.077	-0.099
EX	ee_coverage	0.944	0.940	0.064	0.082	0.100
EX	ex_entropy	0.434	0.494	0.131	0.153	-0.079
EX	ex_prop_away	0.022	0.015	0.825	0.889	-0.006
EX	ee_prop_away	0.030	0.027	0.922	0.922	-0.003

Table 9: MUNDEX: Mann–Whitney tests stratified by annotator role. EE: the explainee's own understanding report; EX: the explainer's judgment of the explainee's understanding. q: BH-corrected within each stratum.

Cond.	Feature	Pos.	Neg.	p	q	r
ec	spk_entropy	0.316	0.379	$< .001$	0.002	-0.078
ec	spk_prop_task	0.728	0.685	$< .001$	0.002	0.078
ec	spk_transitions	0.652	0.773	$< .001$	0.002	-0.073
ec	addr_prop_away	0.049	0.021	$< .001$	0.003	0.029
ec	spk_prop_partner	0.243	0.280	0.001	0.003	-0.070
ec	addr_prop_partner	0.101	0.126	0.004	0.008	-0.051
ec	addr_entropy	0.180	0.215	0.012	0.023	-0.044
ec	addr_transitions	0.330	0.391	0.020	0.033	-0.041
ec	mutual_gaze	0.025	0.035	0.072	0.104	-0.020
ec	spk_coverage	1.000	0.999	0.112	0.146	0.002
ec	addr_coverage	1.000	0.999	0.282	0.324	0.002
ec	spk_prop_away	0.028	0.035	0.299	0.324	-0.008
ec	addr_prop_task	0.850	0.852	0.360	0.360	0.017
nc	spk_prop_partner	0.041	0.054	0.066	0.737	-0.025
nc	addr_transitions	0.079	0.101	0.201	0.737	-0.017
nc	addr_entropy	0.049	0.062	0.203	0.737	-0.017
nc	spk_prop_away	0.028	0.025	0.336	0.737	0.010
nc	spk_prop_task	0.931	0.921	0.343	0.737	0.016
nc	addr_coverage	1.000	1.000	0.408	0.737	-0.001
nc	addr_prop_task	0.880	0.895	0.464	0.737	-0.014
nc	spk_transitions	0.080	0.107	0.548	0.737	-0.007
nc	spk_coverage	1.000	1.000	0.559	0.737	-0.001
nc	spk_entropy	0.039	0.048	0.567	0.737	-0.007
nc	addr_prop_partner	0.049	0.041	0.663	0.751	0.005
nc	addr_prop_away	0.071	0.064	0.694	0.751	0.006
nc	mutual_gaze	0.001	0.000	0.981	0.981	0.000

Table 10: MapTask: association tests stratified by the eye-contact (ec) versus no-eye-contact (nc) condition; q is BH/FDR-adjusted within each condition.

Feature	Coef.	SE	p	q
spk_entropy	-0.108	0.031	$< .001$	0.006
spk_transitions	-0.094	0.029	0.001	0.007
addr_prop_away	0.104	0.040	0.008	0.036
spk_prop_task	0.088	0.042	0.035	0.091
addr_entropy	-0.074	0.036	0.037	0.091
spk_prop_partner	-0.084	0.041	0.042	0.091
addr_transitions	-0.066	0.035	0.061	0.112
mutual_gaze	-0.053	0.031	0.088	0.143
spk_coverage	0.033	0.022	0.141	0.204
addr_prop_partner	-0.060	0.048	0.215	0.279
spk_prop_away	-0.018	0.029	0.539	0.637
addr_prop_task	-0.018	0.043	0.679	0.687
addr_coverage	0.011	0.028	0.687	0.687

Table 11: MapTask: per-feature logistic GEE with dialogue-level clustering. Each row is a separate univariate model. Coef.: standardized log-odds (positive = higher odds of aligned per SD increase); SE: cluster-robust standard error; q: BH-corrected. Features with $q<.05$ are marked † in Table 1.

Feature	Coef.	SE	p	q
ex_prop_partner	-0.351	0.075	$< .001$	$< .001$
ex_prop_task	0.319	0.068	$< .001$	$< .001$
mutual_gaze_explicit	-0.266	0.077	$< .001$	0.002
mutual_gaze_derived	-0.265	0.078	$< .001$	0.002
ee_prop_partner	-0.240	0.082	0.004	0.009
ee_prop_task	0.234	0.081	0.004	0.009
ee_entropy	-0.195	0.070	0.005	0.011
ee_transitions	-0.180	0.069	0.009	0.016
ex_transitions	-0.206	0.082	0.012	0.019
ex_entropy	-0.128	0.077	0.095	0.132
ex_coverage	0.094	0.075	0.209	0.266
ee_coverage	0.074	0.072	0.310	0.361
ex_prop_away	0.054	0.092	0.556	0.599
ee_prop_away	-0.023	0.063	0.709	0.709

Table 12: MUNDEX: per-feature logistic GEE with interaction-level clustering. Columns as in Table 11; positive coefficient = higher odds of UND per SD increase.

Mention	n	Aligned rate	Spk. partner	Spk. task	Spk. entropy	Spk. trans.	Mutual
1	636	0.303	0.230	0.744	0.310	0.643	0.019
2	587	0.595	0.172	0.797	0.215	0.463	0.022
3	518	0.736	0.188	0.791	0.210	0.434	0.018
4+	3403	0.847	0.162	0.808	0.209	0.426	0.016

Table 13: MapTask: how alignment and speaker gaze evolve across repeated mentions of the same landmark. Each row is a mention position (1st, 2nd, . . .); aligned rate is the proportion of mentions at that position that are aligned. Later positions generally show higher alignment and lower partner gaze, though the trend is not strictly monotonic.

Feature	Pre	Post	Δ	p	q
spk_entropy	0.355	0.263	-0.092	0.005	0.044
spk_prop_partner	0.239	0.187	-0.052	0.020	0.079
spk_prop_task	0.755	0.800	0.045	0.052	0.138
spk_transitions	0.720	0.608	-0.111	0.208	0.407
addr_prop_task	0.852	0.874	0.023	0.255	0.407
mutual_gaze	0.030	0.023	-0.007	0.445	0.529
addr_entropy	0.157	0.139	-0.018	0.463	0.529
addr_prop_partner	0.093	0.081	-0.012	0.608	0.608

Table 14: MapTask: gaze at the last non-aligned mention (Pre) vs. the first subsequent aligned mention (Post) of the same concept; same-speaker pairs only (paired Wilcoxon, n=189; q is BH-corrected). A separate mixed-speaker analysis (n=328) was examined as a sensitivity check: the four speaker features change in the same directions and all survive correction, whereas three addressee features change in the opposite direction; no addressee feature survives correction in either analysis.

Mention	Feature	Aligned mean	Non-al. mean	p	q	r
1	spk_transitions	0.554	0.682	0.114	0.200	-0.069
1	spk_prop_partner	0.206	0.240	0.115	0.200	-0.070
1	spk_entropy	0.269	0.327	0.150	0.200	-0.063
1	spk_prop_task	0.763	0.736	0.260	0.260	0.051
2	spk_prop_task	0.836	0.741	0.002	0.010	0.126
2	spk_prop_partner	0.150	0.204	0.059	0.118	-0.076
2	spk_entropy	0.200	0.238	0.166	0.222	-0.054
2	spk_transitions	0.458	0.471	0.352	0.352	-0.036
3	spk_prop_task	0.772	0.845	0.083	0.206	-0.085
3	spk_prop_partner	0.205	0.140	0.103	0.206	0.079
3	spk_transitions	0.449	0.394	0.478	0.638	0.033
3	spk_entropy	0.207	0.218	0.731	0.731	0.016
4+	spk_prop_partner	0.161	0.169	0.351	0.708	-0.021
4+	spk_prop_task	0.809	0.800	0.443	0.708	0.018
4+	spk_transitions	0.422	0.453	0.605	0.708	-0.011
4+	spk_entropy	0.208	0.214	0.708	0.708	-0.008

Table 15: MapTask: aligned vs. non-aligned within each mention-position bucket (Mann–Whitney). Stratifying by position addresses the confound that later mentions are both more aligned and more task-directed. q is BH-corrected within each position bucket (4 tests per bucket). Only task-gaze proportion at second mentions survives correction (q=.01); first-mention contrasts are not significant (aligned rate .303).

Model / feature set	#feat.	Acc.	Macro-F1	$F1_+$	$F1_-$
Majority baseline	–	0.740	0.425	0.851	0.000
LR (controls only)	–	0.514	0.472	0.622	0.322
LR raw proportions	7	0.594	0.522	0.708	0.336
LR structured	13	0.571	0.512	0.682	0.342
+ temporal	55	0.621	0.532	0.736	0.328
+ bigrams	25	0.596	0.521	0.710	0.332
+ coordination	19	0.573	0.511	0.685	0.338
+ ratios	22	0.584	0.516	0.697	0.335
LR all extended	82	0.618	0.527	0.734	0.319

Table 16: MapTask: full prediction results (aligned vs. non-aligned). $F1_+$/$F1_-$ are the aligned/non-aligned classes; "+" rows add one feature group to the LR structured set.

Model / feature set	#feat.	Acc.	Macro-F1	$F1_+$	$F1_-$
Majority baseline	–	0.554	0.356	0.000	0.713
LR (controls only)	–	0.544	0.544	0.550	0.538
LR raw proportions	8	0.564	0.564	0.560	0.568
LR structured	14	0.543	0.543	0.549	0.536
+ temporal	56	0.551	0.550	0.529	0.572
+ bigrams	26	0.523	0.522	0.495	0.548
+ coordination	20	0.551	0.551	0.556	0.546
+ ratios	23	0.545	0.545	0.563	0.526
LR all extended	83	0.527	0.527	0.519	0.534

Table 17: MUNDEX: full prediction results (UND vs. non-UND).

Author Index

Alm, Cecilia O, 89
Asaria, Ali, 72

Berger, Fenna, 150
Berner, Willem, 50
Bisht, Abhishek Singh, 33
Boudreault, Patrick, 89
Bui, Thi Lan Anh, 105
Burenko, Ilya, 59

Cavalcanti, Julio Cesar, 50
Chae, Dong-Kyu, 144
Chaturvedi, Amrita, 33
Cienki, Alan, 150

D'Avirro, Antony, 127
Dang, Trung Duc Anh, 78
Demberg, Vera, 96
Donatelli, Lucia, 150

ElSayed, Salma, 22

Fitch, Allison, 89
Fussell, Susan, 10

Galatolo, Alessio, 160
Gandhi, Deep, 72
Gatt, Albert, 175
Girard, Jeffrey M., 127

Hamilton, Amber, 89
Hanulíková, Adriana, 117
He, Wanqing Psyche, 10
Hilgert, Lukas, 1

Jain, Divij, 96

Kempe, Vera, 22
Kennington, Casey, 105
Kim, Changhyun, 144
Kim, Dahye, 144
Kumar, Rakesh, 33
Kushalnagar, Raja, 89

Lee, Harim, 144
Lee, Sangyool, 144
Li, Nan, 175
Liu, Jia, 96

Manaseryan, Anna, 105
Mohammed, Marwan, 22
Mughal, M. Hamza, 96

Niehues, Jan, 1
Nover, Emily, 89

Paggio, Patrizia, 78
Patterson, Stephanie, 89
Peloquin, Benjamin, 127
Peng, Zhiqing, 78
Poesio, Massimo, 175

Radhe, Tova, 160

Salomone, Tony, 72
Sardhaei, Nasim Mahdinazhad, 40
Seong, Eunseon, 144
Sharifzadeh, Hamid, 40
Skantze, Gabriel, 50
Stebbins, Jillian, 89
Suresh, Varsha, 96

Tolksdorf, Nils Frederik, 117

Uhrig, Peter, 59

Vertino, Jacqueline Romero, 127

Wilson, Anna, 59

Zarrin, Javad, 22
Zheng, Jason, 127
Zygis, Marzena, 40

Åström, Kalle, 50