

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 (Schefflo 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,

	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 MS-DWild (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
<i>Rule-based Baselines</i>									
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)	<u>35.78</u>	<u>35.75</u>	<u>11.91</u>	33.11	<u>31.57</u>	-0.53	32.60	<u>32.37</u>	<u>2.53</u>
Speaker 0	30.49	15.91	0.00	<u>35.97</u>	19.48	<u>0.00</u>	<u>37.85</u>	22.98	0.00
<i>Zero-shot MLLM</i>									
Qwen2.5-Omni-3B	19.70	13.41	0.38	28.20	22.06	<u>0.70</u>	22.99	18.99	-0.37
Qwen3-Omni-30B	<u>30.62</u>	<u>27.59</u>	<u>5.48</u>	<u>30.20</u>	<u>24.28</u>	-2.03	<u>31.64</u>	<u>28.16</u>	<u>0.00</u>
<i>Ground truth text transcript as input for MLLM (Qwen2.5-Omni-3B), no multi-modal information</i>									
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	-	-	-
<i>Text-only MLLM (first transcribe zero-shot, then predict)</i>									
Qwen2.5-Omni-3B	29.07	23.05	-0.62	34.11	<u>29.96</u>	<u>2.42</u>	33.76	28.68	1.06
<i>Fine-tuned MLLM</i>									
Qwen2.5-Omni-3B	<u>48.55</u>	<u>45.48</u>	<u>26.61</u>	37.68	27.77	7.10	36.20	<u>31.06</u>	<u>1.23</u>
LLaVA-Next-Video-7B	30.13	16.06	0.57	35.02	24.57	-0.81	<u>36.34</u>	29.07	-0.40
<i>Fine-tuned MLLM, two-step prompt</i>									
Qwen2.5-Omni-3B	49.26	46.84	26.68	<u>36.15</u>	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, IC-TAI 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, Zheyang 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-3¹
- Qwen3-Omni-30B-A3B-Instruct²
- LLaVA-NeXT-Video-7B-DPO³

• 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://huggingface.co/lmms-lab/LLaVA-NeXT-Video-7B-DPO>

⁵<https://github.com/huggingface/transformers>

⁶<https://github.com/huggingface/transformers>

⁶<https://github.com/huggingface/transformers>

- 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
<i>Fine-tuned MLLM (Qwen2.5-Omni-3B), AMI + MSDWild, vary context</i>									
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
<i>Rule-based Baselines, evaluate different turns</i>									
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
<i>Fine-tuned MLLM (Qwen2.5-Omni-3B), evaluate different turns</i>									
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
<i>Fine-tuned MLLM (Qwen2.5-Omni-3B), different training mix</i>									
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.