

Mind the Motions 🧠: Benchmarking Theory-of-Mind in Everyday Body Language

Seungbeen Lee⁺, Jinhong Jeong⁺, Donghyun Kim⁺, Yejin Son⁺, Youngjae Yu[◇]

⁺Yonsei University, [◇]Seoul National University

seungblee@yonsei.ac.kr

Abstract

Our ability to interpret others' mental states through nonverbal cues (NVCs) is fundamental to our survival and social cohesion. While existing Theory of Mind (ToM) benchmarks have primarily focused on false-belief tasks and reasoning with asymmetric information, they overlook other mental states beyond belief and the rich tapestry of human nonverbal communication. We present 🧠MOTION2MIND¹, a framework for evaluating the ToM capabilities of machines in interpreting NVCs. Leveraging an expert-curated body-language reference as a knowledge base, we build 🧠MOTION2MIND, a carefully curated video dataset with fine-grained nonverbal cue annotations paired with manually verified dictionary-grounded interpretations. It encompasses 222 types of nonverbal cues and 397 mind states. Our evaluation reveals that current AI systems struggle with NVC interpretation, exhibiting not only a substantial performance gap in *Detection*, as well as patterns of over-interpretation in *Explanation* compared to human annotators.

1 Introduction

Understanding **others' mental states through visual cues** is fundamental to human social interaction and intelligence (Tomasello et al., 2005). We naturally infer emotions from facial expressions (Barrett et al., 2011), intentions from behaviors (Becchio et al., 2018), and social status from appearances (Freeman and Ambady, 2011). As artificial intelligence systems become increasingly integrated into our daily lives—from virtual assistants to social robots (Mathur et al., 2024)—their ability to interpret these NVCs becomes crucial for meaningful human-AI interaction. Fields such as affective computing and social robotics heavily rely on human nonverbal interaction (Picard et al., 2001; Spezialetti et al., 2020).

However, implementing robust nonverbal Theory of Mind (ToM) faces two critical challenges. First, current evaluation systems create a diagnostic challenge due to their end-to-end mapping which link raw visual cue to final mind states. It is difficult to isolate whether the system failed in visual detection, reasonable attribution to the cue, or holistic contextual thinking. Second, nonverbal behaviors exhibit severe contextual ambiguity. The same cue maps to different interpretations depending on the situation, and while humans share common foundational patterns, individual perception still carries subjectivity to recognize and interpret the cue. Existing datasets often overlook these challenges, lacking localized tracking or structured error diagnostics.

We introduce 🧠MOTION2MIND, a human-verified evaluation benchmark to test NVC understanding of Large language model (LLM). First, as illustrated in Figure 1, we separate nonverbal cue understanding into a three-stage pipeline, pairing our video dataset (V) with fine-grained cue annotations (C) and contextually grounded mind-state annotations (M). This multi-layer design allows us to pinpoint the exact computational bottleneck within the perception-to-reasoning flow. Second, we construct a meticulously audited dataset encompassing 1,022 instances that cover 222 cue types, all strictly aligned with a consolidated body-language reference dictionary and human inspection. This carefully controlled small-scale testbed serves as a reliable baseline to evaluate over-interpretation and under-interpretation, providing a solid foundation for scaling up future nonverbal ToM research.

Empirically, our evaluation reveals a consistent human-AI performance gap especially in *Detection* and *Explanation* task. LLMs exhibit degradation in contextual inference, particularly in the presence of invalid cues. Notably, we observe a systematic tendency toward over-interpretation, where models assign confident psychological explanations even

¹ 🧠 seunbite/motion2mind

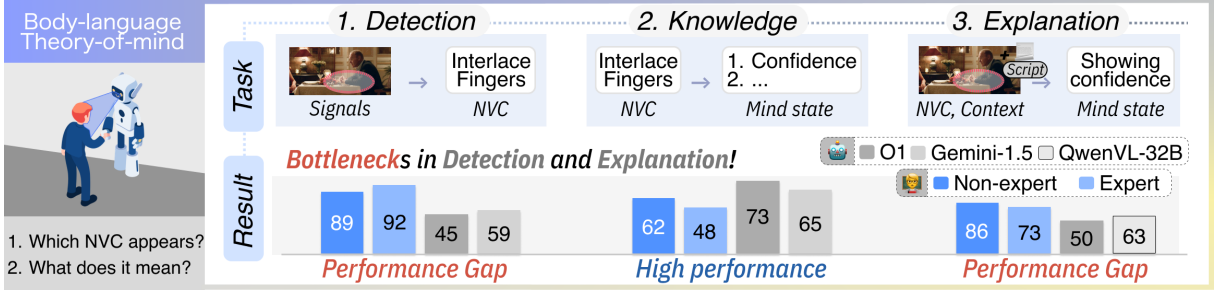


Figure 1: The three-stage framework for nonverbal cue understanding: (1) **Detection** ($\mathcal{V} \rightarrow \mathcal{C}$): identifying NVC from movements; (2) **Knowledge** ($\mathcal{C} \rightarrow \{\mathcal{M}\}$): context-free NVC interpretation; and (3) **Explanation** ($(\mathcal{C}, \mathcal{S}) \rightarrow \mathcal{M}^*$): contextual reasoning for mind-states related to the NVC. While LLMs match humans in textbook Knowledge, MOTION2MIND exposes a substantial performance gap in both Detection and Explanation.

for non-informative behaviors, highlighting a key limitation in current multimodal reasoning systems. Our key contributions are:

1. **A Formal Three-Stage Framework for Non-verbal ToM:** As illustrated in Figure 1, we propose a structured framework that decomposes nonverbal mentalization into three computationally trackable phases: *Detection*, *Knowledge*, and *Explanation*.
2. **MOTION2MIND with Contextually Invalid Cues:** We construct a multimodal benchmark paired with 1,022 manually verified instances. Crucially, we introduce *invalid cues* (visually salient but no latent psychological meaning) as a strict negative control to diagnose the over-interpretation tendencies of current AI systems.
3. **Grounded Multi-Modal Evaluation and Human Baselines:** We formulate five comprehensive task types across distinct input modalities to pinpoint execution bottlenecks. Our comparative analyses against both non-experts and clinical psychologists map a significant human-AI performance gap.

2 Understanding Nonverbal Cue

Many psychological studies divide the mentalization process into successive stages (Fonagy, 2006). To evaluate the performance of NVC understanding, we break down the process where external stimuli are transformed into mental-state inferences.

2.1 Detection / Perception ($f : \mathcal{V} \rightarrow \mathcal{C}$)

Detection converts raw multimodal signals ($\mathcal{V}$) into discrete nonverbal cue ($\mathcal{C}$) recognition. Accurate detection is a prerequisite for downstream

inference. Key challenges include the fact that an agent’s capability to recognize them varies across different individuals and is influenced by dynamic environmental factors (*e.g.* context, camera angles, occlusions, and background noise).

2.2 Knowledge ($g : \mathcal{C} \rightarrow \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$)

The knowledge component maps each detected cue ($\mathcal{C}$) to a set of *plausible* psychological interpretations ($\{\mathcal{M}_1, \mathcal{M}_2, \dots\}$). A central challenge in non-verbal communication analysis is the absence of clear ground truth and high context-dependency, allowing multiple competing hypotheses for a single observed behavior. For this reason, we adopt an established body-language reference (Navarro, 2018) as a representative *knowledge base* rather than an absolute psychological truth. This framework provides an operationalizable inventory of cue-meaning associations. While we acknowledge that body language interpretation is culturally dynamic and scientifically non-monolithic, using a single consistent reference prevents introducing conflicting taxonomies during the annotation baseline stage. With this knowledge base, we evaluate how well models align with a consistent and human-documented prior under controlled conditions.

2.3 Explanation ($h : (\mathcal{C}, \mathcal{S}) \rightarrow \mathcal{M}^*$)

Explanation takes the candidate interpretations from the knowledge component and combines them with contextual information ($\mathcal{S}$) to yield a final mental-state hypothesis ($\mathcal{M}^*$) (*e.g.* surprised). This stage addresses the inherent ambiguity of nonverbal behavior by leveraging environmental cues.

Terminology. We use *nonverbal cue* (NVC) for observable gestures, poses, or vocal prosody, and *mind state* for the latent psychological interpreta-

	Cue → Explanation	Explanation → Cue
Prompt	Given a nonverbal cue, please choose the most plausible explanation from the options. <i>‘Arm crossing’</i>	Given the explanation of a nonverbal cue, please provide a plausible nonverbal cue from the options. <i>‘Feeling insecure or threatened’</i>
Options	0: Enthusiastic celebration 1: Drive to emphasize key statements 2: Feeling insecure or threatened 3: Wanting to connect or belong	0: Arm crossing 1: Elation triumph displays 2: Elbow flexing 3: Hugging

Table 1: Illustration of our bidirectional NVC evaluation tasks. In §3, we implement two-sided tasks: Cue to Explanation and Explanation to Cue.

tion (emotion, attitude, or intention).

3 Knowledge: Body-language understanding Without Context

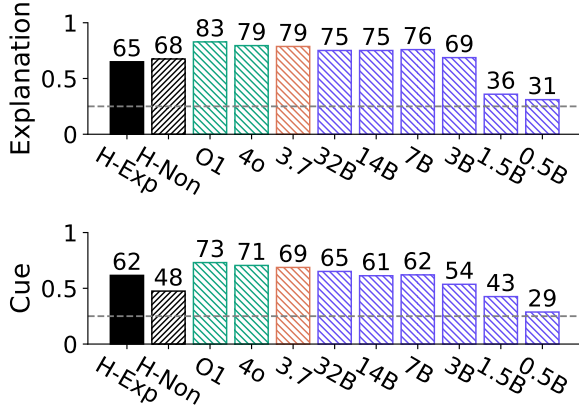


Figure 2: NVC knowledge scores of intelligent LLMs — GPT (green), Claude (orange), Qwen2.5-Instruct (purple) — tested on the NVC dictionary. LLMs manifest structured knowledge more accurate than psychological experts.

We test how prior knowledge of state-of-the-art LLMs (GPT, Claude, Qwen2.5-Instruct) aligns with the structured NVC dictionary.

3.1 Methodology

Test Set Dictionary covers 407 NVCs and their possible (multiple) mind state interpretations. To process this, we structure the consolidated explanation paragraph into n different semantic units (*e.g.* Fatigue, Stressed, Interested) using GPT-o1.

Tasks As shown in Table 1, we design two task types to measure NVC proficiency.

1. **Cue → Explanation (Understanding):** Models select the most plausible interpretation of a given nonverbal cue.

2. **Explanation → Cue (Generation):** Models generate a matching cue from an explanation.

Multi-choice with Mutually Exclusive Negatives

We simplify the evaluation into a multiple-choice QA format. Constructing the options, distractors are randomly sampled from the entire pool of explanations. To account for the multi-answer nature of nonverbal communication ($\{M_1, M_2, \dots\}$), we utilize cosine similarity which filter out candidates with cosine distance of 0.8 or higher from *all* valid explanations. Then we randomly sample three distractors, and finally shuffle the position of the correct answer to eliminate positioning bias.²

Human Baselines Performance is measured against two human groups: (1) four experts: psychologists with counseling certificates and (2) five non-experts: graduate students with no psychology expertise. This dual baseline highlights gaps between LLMs and human understanding. More details about annotators are in Appendix C.

3.2 Results

LLMs Align in Textbook Knowledge. Figure 2 show that top-tier LLMs achieve high alignment with the structured dictionary. However, we note that this high text-only accuracy may be partly inflated by data contamination, as these web-scale models likely encountered the reference text during pre-training. Interestingly, non-experts marginally outperform experts in the context-free mapping (68% vs 65%). This is because context-free setups force non-experts to simply choose the most representative, commonsense definition. In contrast, clinical psychologists naturally entertain a wider,

²We use OpenAI’s ‘text-embedding-3-small’ for computing semantic embeddings.

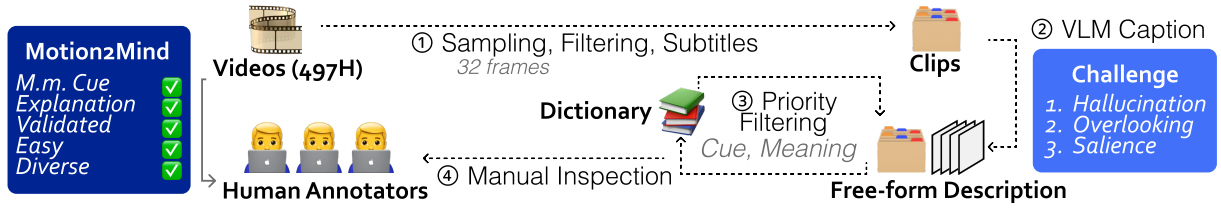


Figure 3: Overview of the MOTION2MIND construction pipeline. We collect 497 hours of diverse YouTube videos, sample 32-frame clips, and generate initial captions using Qwen2.5-32B-VL-Instruct. Captions are filtered via a body-language dictionary to isolate interpretable cues. Finally, human experts manually inspect and refine the interpretations based on contextual grounding to ensure semantic saliency.

multi-layered spectrum of hidden diagnostic possibilities for an isolated gesture, which may lead to higher answer variance and lower accuracy when stripped of situational context.

Large models do better in both tasks. Across evaluated models, we observe a clear scaling effect where larger variants (*e.g.* GPT-o1, Qwen2.5-32B) consistently navigate these structured associations better than smaller ones. More notably, all models score higher on the Cue $\rightarrow$ Explanation (*Understanding*) task than on the Explanation $\rightarrow$ Cue (*Generation*) task. This gap shows that while AI systems are good at recognizing the meaning of observed behavior, they struggle to map abstract psychological states back to the corresponding physical actions.

4 MOTION2MIND

Figure 3 provides an overview of pipeline to make MOTION2MIND, a carefully curated video dataset designed to test body language understanding within contexts. It features (1) video clips sourced from YouTube content; (2) fine-grained motion annotations on short 4-second segments, each paired with psychological interpretations grounded in the full context; and (3) high-quality annotations validated by human psychologists, achieving average 92% accuracy.

4.1 Video Collection

Diverse Real-World Sources We source diverse videos from six high-subscriber YouTube channels spanning film, television, and reality genres. Using the YT-DLP framework³, we collected approximately 4,730 unique clips of total 497.92 hours.

Clip Sampling We randomly sample short 4-second segments rather than exhaustively processing entire videos. Each 4-second clip is extracted

at 8 frames per second (fps), a rate chosen based on empirical tests balancing visual informativeness and computational efficiency.

Filtering To ensure the presence of NVCs, we filter out clips that either (1) lack human presence or (2) inconsistent frame-wise people detection which means scene transitions. We apply YOLOv8 (Varghese and Sambath, 2024) to detect human presence and track consistency across frames.

Subtitles We extract spoken dialogue with Whisper-large-v3, segment speakers with NVIDIA NeMo, detect body parts with MediaPipe, and analyze speech with Praat.⁴

4.2 Nonverbal Cue

We annotate both visual and vocal nonverbal cues using a hybrid approach of automatic pipelines and detailed human inspection.

4.2.1 Visual Cue

Challenge We initially test Qwen2.5-32B-VL-Instruct for free-form captioning of short video clips. Despite the promising abilities, it introduces several common issues: (1) *Hallucination* and *Omission*: Describing not observed cues or overlooking observed cues. (2) *Misalignment with Human Saliency*: The described cue is present but not the most overt for the human.

Solution 1: Body Part Detection and Prompt Conditioning We apply MediaPipe to identify visible body parts (*e.g.*, face, arms, hands, torso) in each clip. This serves hallucination filtering to discard captions from non-detected body parts

³<https://github.com/yt-dlp/yt-dlp>

⁴Whisper: <https://github.com/openai/whisper>, <https://huggingface.co/openai/whisper-large-v3>; NeMo: <https://github.com/NVIDIA/NeMo>; MediaPipe: <https://developers.google.com/mediapipe>; Praat: <https://www.praat.org/>.

Dataset	Items	Mods	Cue.	Vocal.	Ambiguity	Invalid	Source
 MOTION2MIND	1,022	V + A + T	✓	✓	✓	✓	Movie, Sitcom, Reality
SOCIAL GENOME (Mathur et al., 2025)	272	V + A + T	✓	✓	✗	✗	YouTube
MMToM-QA (Jin et al., 2024)	1.0k	V + A + T	✓	✗	✗	✗	Simulation
MuMA-ToM (Shi et al., 2025)	1.6k	V + A + T	✓	✓	✗	✗	Simulation
R3-VQA (Niu et al., 2025)	5.1k	V + A + T	✓	✓	✓	✗	YouTube / TV / Movie
VEATIC (Ren et al., 2024)	124	V + A	✗	✓	✓	✗	Mixed clips
MovieGraphs (Vicol et al., 2018)	7.6k	V + T	✓	✗	✓	✗	Movies
VEGAS (Li et al., 2025)	1.2k	V + T	✓	✗	✗	✗	YouTube

Table 2: We introduce  MOTION2MIND, the first multimodal dataset with fine-grained motion annotations and validated psychological explanations. Mods denotes modality in data. V = vision, A = audio, T = text. Cue. denotes annotation of fine-grained nonverbal cue. Vocal. = annotation of vocal nonverbal cue. Ambiguity indicates whether the target mental states encompass subjective emotions or implicit desires that cannot be definitively deduced from explicit context clues alone. Items denotes the primary annotated unit reported by each dataset (e.g., videos, clips, or QA instances).

and focused prompting to produce more specific descriptions.

Solution 2: Dictionary-Guided Priority Filtering We convert the free-form captions into structured JSON format using GPT-4o-mini. Each entry includes the detected *cue*, *actor*, *body parts*, and an *explanation* if specified. To narrow the candidate set and sort with priority, we filter out any cues not found in our reference body language dictionary (has extremely low semantic similarity with any dictionary entity). This step eliminates subjective or overly creative outputs and narrows the candidate set for human review.

Solution 3: Final Human Inspection Remaining annotations are manually reviewed by the authors. Three criteria are used: (1) *Appearance*: Is the described cue visibly present in the clip? (2) *Salience*: Is it the most psychologically relevant cue in the scene? (3) *Diversity*: Are the numbers of NVC balanced? This step ensures that annotations are both accurate and balanced.

4.2.2 Vocal Cue

Our vocal cue annotation pipeline identifies three primary vocal cues: Speaking rate, Pitch, and Silence duration.

Speaking Rate, Pitch, Long Pause To automatically label the fast speaking, we measure words per minute (WPM) within each segment, dynamically applying the mean and standard deviation for speaker, then label [Fast] when normalized WPM exceeds 1.5. Similarly, using automatic package such as Praat, we annotate [HIGH_PITCH] when normalized pitch surpasses 1.25. Finally, with We-

bRTC VAD, Segments with a silence duration exceeding 600 ms and accounting for over 5% of the total segment length are labeled as [LONG_PAUSE].

4.3 Interpretations

Challenges Interpreting NVCs presents three major challenges. (1) *Ambiguity*: Many cues have multiple possible meanings or no clear interpretation depending on contexts; (2) *Subjectivity*: Perceptions vary between observers; (3) *Over-interpretation*: Automatic pipeline such as VLMs tend to assign meaning to every cue.

Solution 1: Dictionary-Constrained Interpretation To mitigate *Ambiguity*, we constrain all NVC interpretations to our predefined dictionary containing 2,050 possible psychological explanations. Based on exact or inclusive string matching, our ‘valid cues’ and their explanations directly align with the dictionary phraseology.

Solution 2: Human-Guided Labeling and Invalid Cases To address *Subjectivity* and evaluate model robustness against over-interpretation, we partition the  MOTION2MIND into a binary structural taxonomy comprising 1,022 total instances:

- **Valid Cues (852 items, 83.4%)**: Instances where the observed visual or vocal behavior carries a distinct, contextually supported psychological mind-state specified in dataset.
- **Invalid Cues (170 items, 16.6%)**: Instances where a visually salient movement occurs (e.g. “fingers close together” or simple shifting) but carries no consensus-supported meaning or sufficient contextual grounding within that specific scene.

Model	Open	Input	ToM Method	Detection		Cue	Explanation			Prediction
				MCQ	Binary	Accuracy	Total	Valid	Invalid	MCQ
<i>Expert</i>	–	–	–	–	89.0	–	81.3	76.3	86.3	90.0
<i>Non-expert</i>	–	–	–	–	92.0	–	69.3	63.3	73.3	83.3
GPT-o1	✗	V, T, (A)	✗	64.3	45.0	40.6	62.5	64.9	50.6	95.7
GPT-4o	✗	V, T, (A)	✗	64.3	45.4	41.1	62.3	64.9	49.4	67.9
Gemini-Flash-1.5	✗	V, T, A	✗	67.6	59.2	64.9	46.2	65.2	63.5	73.8
Qwen 2.5-32B	✓	V, T, (A)	✗	34.9	77.0	42.9	42.5	65.5	30.0	86.7
Qwen 2.5-7B	✓	V, T, (A)	✗	31.9	53.1	41.0	41.0	65.1	29.6	78.8
Qwen 2.5-3B	✓	V, T, (A)	✗	58.8	54.0	44.2	47.8	57.3	0.0	25.7
InternVL3-8B	✓	V, T, (A)	✗	68.0	78.0	54.0	59.9	66.0	29.5	81.5
InternVL3-2B	✓	V, T, (A)	✗	67.0	95.6	49.6	43.8	51.3	6.5	68.9
Qwen 2.5-32B	✓	V, T, (A)	Wilf et al. (2024)	-	-	59.2	61.8	65.6	40.0	67.0
Qwen 2.5-7B	✓	V, T, (A)	Sclar et al. (2023)	-	-	58.3	51.4	56.0	25.3	64.9

Table 3: Performance of VLMs on MOTION2MIND across five tasks: (1) *Detection* MCQ ($\mathcal{V} \rightarrow \mathcal{C}$), (2) *Detection* Binary ($(\mathcal{V}, c) \rightarrow \{0, 1\}$), (3) *Cue* ($(\mathcal{V}_{\text{pre}}, \mathcal{S}) \rightarrow c^*$), (4) *Explanation* ($(\mathcal{V}, \mathcal{S}, c) \rightarrow m^*$), and (5) *Prediction* ($(\mathcal{V}, \mathcal{S}, c) \rightarrow u^*$). VLMs consistently underperform human baselines. The random baseline is 25% for all multiple-choice tasks except Detection–Binary (50%). Input modalities include Visual (V), Text (T), and Audio (A).

To ensure the highest fidelity, all annotations underwent rigorous cross-checking by a second human annotator.

5 Test VLMs

We test current VLMs’ performance on our MOTION2MIND benchmark. Specifically, we test GPT o1 and 4o, Qwen2.5-VL 32B to 3B, and InternVL 8B to 2B.

Multi-choice We formulate this task as a multiple-choice question. For the Explanation and Cue tasks, options are drawn from the pool of cue–explanation pairs in the dictionary; for Prediction, from the pool of utterances in scripts. As in §3, we filter out options with high semantic similarity to the ground-truth answer to guarantee discriminative evaluation.

5.1 Input Modality

Visual (Frames, NVC) We provide a sequence of video frames as visual input, representing a 4-second clip containing the target NVC. To stay within the model’s visual-token limit, we down-sample each clip to a maximum of 32 frames, with a minimum frame resolution of 64 pixels.

Textual (Script, Context) For the *Cue*, *Explanation*, and *Prediction* tasks, we supply up to 60 seconds of dialogue script as a textual context. Vocal events (*e.g.* sighs, laughter) are annotated inline to preserve prosodic information.

Human annotator input Human annotators are presented with the same visual clips and textual context as models, without access to raw audio. However, replay is permitted given the limitations of short-term memory. Due to annotation cost, human evaluation is conducted on a representative subset of tasks (Detection–binary, Explanation, and Prediction), which we found sufficient to capture the human–model performance gap.

5.2 Task Definition

Detection ($\mathcal{V} \rightarrow \mathcal{C}$ or $\{0, 1\}$) The goal is to identify which nonverbal cue (C) appears in the given visual input (V). We design two formats: (1) MCQ, where the model selects the correct cue from multiple distractors, and (2) Binary, where the frames and a candidate cue is provided, and the model chooses between ‘Appears.’ and ‘Does not appear’.

Cue (Generation) ($\{\mathcal{V}_{\text{pre}}, \mathcal{S}\} \rightarrow C$) The task is to infer the most plausible nonverbal cue in the blank. We also provide the preceding 4-second video chunk ($\mathcal{V}_{\text{pre}}$) to supply relevant visual context while avoiding spoilers.

Explanation ($\{\mathcal{V}, \mathcal{S}, C\} \rightarrow M$) Similarly given a short video clip and its aligned script, and also the specified nonverbal cue, the model is asked to infer the most likely underlying psychological state (M). This task evaluates the model’s ability to interpret the meaning of observed behavior.

Prediction (Next Utterance) ($\{\mathcal{V}, \mathcal{S}, C\} \rightarrow U^*$) This task provides both the visual clip and its sur-

rounding script, with a blank for the next line of dialogue following the nonverbal cue (U). This serves as a proxy for its ability to reason about mental state transitions in context.

5.3 Results

Explanation, Detection: Clear Human-AI Gap

As shown in Table 3, even non-expert humans outperform all tested models on key tasks. Experts reach over 80% accuracy on Explanation and 90% on Prediction. Although there is a strong scaling effect, the best VLM shows clearly lower capability (o1: 45.0) than human (Experts: 89.0).

Explanation: Struggles with Invalid Cues Models consistently connect ‘invalid’ nonverbal cue with certain meaning. Most models show a 30–40 point gap between valid and invalid Explanation accuracy, suggesting a tendency to over-interpret. We further analyze in §5.5.

Theory of Mind Modules Yield Limited Gains

While models with ToM modules (Wilf et al., 2024; Sclar et al., 2023; Jin et al., 2024) show modest gains (e.g. 32B, Explanation 64.9 in certain tasks, they do not close the gap with human performance. For example, Qwen 2.5-32B with ToM achieves only a slight improvement over GPT-4o without ToM in Explanation (65.5% vs 64.9%). Cue prediction remains particularly challenging across all configurations. Full results using ToM Baselines are in Appendix D, and we adopt the best approach for two models in Table 3.

Detection: Binary vs. Multi-choice Binary Detection is generally easier than the MCQ variant, likely due to lower ambiguity in answer choices. However, some models (e.g., InternVL3-2B at 95.6%) show unrealistically high scores, likely due to overfitting to the default ‘Appears’ label in the binary setting.

Explanation, Prediction: Larger Models Excel in Contextual Tasks

These tasks require nuanced, context-dependent reasoning, and this highlights the benefits of larger models. For example, Qwen 2.5-32B and GPT-o1 outperform smaller models by over 20 points in Prediction.

5.4 Modality Ablation

To disentangle the contribution of each modality, we evaluate Qwen2.5-VL-32B under three input conditions: (1) *Full* (visual frames + script),

Task	Qwen2.5-VL-7B			Qwen2.5-VL-32B		
	Full	Vis.	Text	Full	Vis.	Text
Cue	41.0	33.0	41.4	42.9	33.6	40.4
Explanation	41.0	29.3	40.6	42.5	30.0	42.8
Detection	31.9	31.6	40.7	34.9	34.7	43.8
Detection (binary)	53.1	52.5	33.2	77.0	77.1	30.2
Next prediction	78.8	30.5	76.7	86.7	29.1	73.1

Table 4: Modality ablation (accuracy %). Bold is the best of Full / Vision-only / Text-only within each model.

(2) *Text-only* (script without video), and (3) *Vision-only* (frames without textual context).

Performance Degradation under Visual Grounding

Table 4 shows that no single modality is uniformly best across tasks for either 7B or 32B, including the Full setting that receives both video and script. Comparing Vision-only with Text-only, textual context more often helps models recover meaning (Cue, Explanation, and 4-way Detection). By contrast, on binary Detection the visual stream is already highly informative—Vision-only matches or slightly exceeds Full, while Text-only collapses—so seeing the cue matters more than reading about it. Next-prediction still favors Full over either unimodal input, indicating that the two sources are complementary rather than redundant.

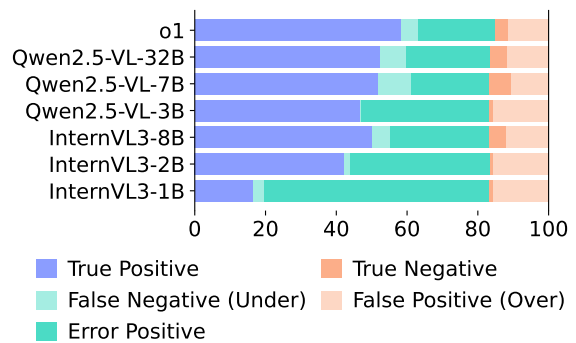


Figure 4: Stacked bar plots of *Explanation* task answers. Small models shows low precision (over-interpret) compared to larger models.

5.5 Over- vs. Under-interpretation

In *Explanation* task, we categorize the combination of (‘valid’ / ‘invalid’) type of model answers and ground-truth to study overinterpretation and underinterpretation of LLM answers. Underinterpretation is structurized as False Negative (model: valid, groundtruth: invalid) and Overinterpretation as False Positive (model: invalid, groundtruth: valid).



Figure 5: Examples of erroneous inferences by the GPT-O1 model in Detection-Binary and explanation tasks. The first row illustrates the example which model doesn’t recognize the given cue (*e.g.* Smile, Neck touching). The second row presents misinterpretations, where benign or contextually ambiguous cues are incorrectly assigned psychological meanings (F: False explanation, T: True explanation).

Predominance of Over-Interpretation In Figure 4, despite a ground-truth skew toward valid explanations, over-interpretation (False Positives) far outnumbers under-interpretation (False Negatives). Models rarely confuse a valid cue for an invalid one. As model size decreases, the proportion of ‘Error Positives’ (EP)—instances where a model labels a cue as valid but assigns the wrong explanation—rises sharply.

5.6 Qualitative results

Figure 5 shows representative cases where the O1 model produces incorrect inferences in Detection-Binary and Explanation tasks. In Detection-Binary task, the model misidentifies even clear cues such as ‘smiling’ and ‘gesturing while speaking’. In the explanation tasks, the model demonstrates a tendency to over-interpret benign cues as indicative of psychological states.

6 Related Work

Theory of Mind Benchmarks Early AI ToM benchmarks largely mirror developmental false-belief tests in text form (Le et al., 2019; Kim et al., 2023; Li et al., 2023; Amirizani et al., 2024), some papers encompassing visual cues as input (Jin et al., 2024; Etesam et al., 2024; Ma et al., 2023) evaluating models’ ability to distinguish asymmetric information in templated stories. Recent efforts expand ToM assessments to broader men-

tal states—emotions, intentions, desires, beliefs, knowledges, percepts—and incorporate visual context (Wang et al., 2025; Ma et al., 2023; Duan et al., 2022; Fan et al., 2021; Mao et al., 2024; Bortoletto et al., 2024) utilizing agent behavior or navigation as the inferred cue. 🧠MOTION2MIND deals with nuanced and detailed body language sourced from a structured NVC dictionary.

Video-Based Social Reasoning NVC datasets are built in video understanding domain to classify the appropriate emotion state or social relation of the character in the video (Luo et al., 2020; Liu et al., 2021; Huang et al., 2021; Zadeh et al., 2019; Chen et al., 2024; Vicol et al., 2018; Lei et al., 2020). Social Genome (Mathur et al., 2025) introduces 272 videos paired with 1,486 human-annotated reasoning traces. Social Genome deals with multimodal social-reasoning chains with diverse information type, but our 🧠MOTION2MIND focuses on visual information in the domain of body language.

7 Conclusion

Our study presents a comprehensive evaluation framework, 🧠MOTION2MIND, for assessing AI systems’ capacity to interpret nonverbal cues (NVCs) in real-world, multimodal contexts, revealing substantial gaps between human and machine performance. Their performance degrades when faced with contextual ambiguity and nuanced so-

cial cues (Invalid). State-of-the-art models such as GPT-4o and Qwen2.5-VL fail to consistently integrate visual and textual modalities, as evidenced by inconsistent performance in combined *Detection* and *Explanation* tasks.

8 Limitations

Single-Source Dependency Our framework relies on a single body-language reference as a knowledge base. While this enables structured evaluation, it may introduce biases specific to the source, including stylistic or cultural biases. Future work should incorporate multiple references and empirically study agreement and divergence across sources.

Dataset Scale vs. Precision While our dataset comprises a meticulously audited 1,022 instances with high-density annotations, the human evaluation baseline is bounded by a smaller sample size (4 experts and 5 non-experts). While these numbers provide a clear indication of a performance gap, future extensions will focus on computing broader confidence intervals and inter-annotator agreement metrics across a larger cohort to reinforce statistical robustness against diverse edge cases.

Cultural Variability Nonverbal meanings vary across cultures, social roles, and interaction settings. Although the reference includes some non-Western cues (*e.g.* Namaste, Gaze hierarchy, and Kowtow) and some culture-neutral cues (*e.g.* Proximity, Smiling, and Leaning) real-world interpretation varies widely across societies. Understanding and correcting for cultural bias in video-language models is therefore an important and independent direction for future work.

9 Ethical Considerations

Privacy and Consent While our video dataset uses publicly available YouTube clips, the broader application of NVC understanding raises important privacy concerns. The ability to automatically interpret body language and emotional states could enable surveillance systems that infringe on personal privacy. Future deployments of such technology should carefully consider consent mechanisms and privacy protections, particularly in public spaces or workplace environments.

Potential for Misuse and Manipulation Advanced understanding of NVCs could be exploited for manipulation or deception. Systems capable of

interpreting subtle behavioral signals might be misused for psychological profiling, social engineering, or targeted influence campaigns. Additionally, the technology could be used to develop more sophisticated deepfake systems that incorporate realistic nonverbal behaviors, further complicating issues of digital authenticity and trust.

Bias and Cultural Sensitivity Our framework, despite efforts to be comprehensive, may contain inherent biases in how it interprets and validates NVCs across different cultural contexts. Reliance on Western-centric sources for body language interpretation could lead to misinterpretation or oversimplification of culturally specific gestures and expressions. Furthermore, the use of movie clips as a data source may perpetuate certain cultural stereotypes or biases in the portrayal and interpretation of emotional states.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No. RS-2026-25522885, Development of a World Foundation Model for Training and Development of Physical AI Systems, No. RS-2026-25512061, Development of Large Action Model-Based Autonomous Digital Twin Operation Technology for Proactive Problem Solving), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2024-00354218, No. RS-2026-25561904), and the Technology Innovation Program (RS-2025-25456760, Development of a humanoid robot specialized in chemical processes based on AI foundation model) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program). We express special thanks to KAIT GPU project. This work was supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation. We thank Harim, Dayi, and Minhee for their help with data annotation and validation.

References

- Maryam Amirizani, Elias Martin, Maryna Sivachenko, Afra Jahanbakhsh Mashhadi, and Chirag Shah. 2024. [Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses](#). *ArXiv*, abs/2406.05659.
- Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current directions in psychological science*, 20(5):286–290.
- Cristina Becchio, Atesh Koul, Caterina Ansuini, Cesare Bertone, and Andrea Cavallo. 2018. Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of life reviews*, 24:67–80.
- Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024. Explicit modelling of theory of mind for belief prediction in non-verbal social interactions. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain—including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pages 866–873. SAGE Publications Pvt. Ltd 1 Oliver’s Yard, 55 City Road, London, EC1Y 1SP.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *CoRR*.
- Jiafei Duan, Samson Yu, Nicholas Tan, Li Yi, and Cheston Tan. 2022. Boss: A benchmark for human belief prediction in object-context scenarios. *arXiv preprint arXiv:2206.10665*.
- Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Emotional theory of mind: Bridging fast visual processing with slow linguistic reasoning. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 10–18. IEEE.
- Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. 2021. Learning triadic belief dynamics in nonverbal communication from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7312–7321.
- Peter Fonagy. 2006. The mentalization-focused approach to social development.
- Jonathan B Freeman and Nalini Ambady. 2011. A dynamic interactive theory of person construal. *Psychological review*, 118(2):247.
- Yibo Huang, Hongqian Wen, Linbo Qing, Rulong Jin, and Leiming Xiao. 2021. Emotion recognition based on body and context fusion in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3609–3617.
- Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877.
- Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8769–8784.
- Hao Li, Hao Fei, Zechao Hu, Zhengwei Yang, and Zheng Wang. 2025. Vegas: Towards visually explainable and grounded artificial social intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4707–4715.
- Huaoli, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192.
- Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10631–10642.

- Yu Luo, Jianbo Ye, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. 2020. Ar-bee: Towards automated recognition of bodily expression of emotion in the wild. *International journal of computer vision*, 128(1):1–25.
- Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1011–1031.
- Yuanyuan Mao, Xin Lin, Qin Ni, and Liang He. 2024. Bdiqa: A new dataset for video question answering to explore cognitive reasoning through theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 583–591.
- 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.
- Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24879–24902.
- J. Navarro. 2018. *The Dictionary of Body Language: A Field Guide to Human Behavior*. HarperCollins.
- Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. 2025. $r^3 - vqa$: “read the room” by video social reasoning. Preprint, arXiv:2505.04147.
- Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191.
- Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. 2024. Veatic: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4467–4477.
- Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models’ (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.
- Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 1510–1519.
- Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279.
- Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691.
- Rejin Varghese and M Sambath. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE.
- Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590.
- Qiaosi Wang, Xuhui Zhou, Maarten Sap, Jodi Forlizzi, and Hong Shen. 2025. Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective. *arXiv preprint arXiv:2504.10839*.
- Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models’ theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.
- 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.

A Dataset Construction Details

A.1 Video Sources and Sampling Strategy

- **Source Channels:** Sitcoms (Clipzone Sitcoms, *The Office*, *Friends*), Movies (Lionsgate, JoBlo), Reality Shows (*Keeping Up with the Kardashians*).
- **Sampling Protocol:** We sampled up to 40 clips (each 4 seconds long) per video.

A.2 Filtering and Preprocessing

- **Cue Filtering:** We applied semantic and lexical matching using Sentence-BERT embeddings against the dictionary.
- **Rejection Criteria:** Non-human content, poor visibility, occlusion, and rapid cuts were removed.

A.3 Human Inspection Process

- **Manual Review:** 24.3% stratified sample was manually inspected.
- **Results:** The relatively low pass rate reflects the difficulty of identifying clips with sufficiently clear, contextually grounded, and socially meaningful nonverbal behaviour under our annotation criteria; Inter-rater agreement (Cohen's κ) = 0.79 on 100 items.

B Cue Dictionary

B.1 Statistics

- **19 anatomical categories** (Head, Eyes, Eyebrows, Mouth, Hands, Torso, Feet, etc.).
- An average of **21.4 cues per category**, yielding a total of **407 NVCs**.
- For each cue, **5.03 psychological explanations** on average, spanning Knowledge, Beliefs, Percepts, Desires, and Emotions (Table 5).

B.2 Cultural Limitations and Future Work

- Rooted in Western sources; includes some non-Western cues (e.g. *Namaste*, *kowtow*) and global cues (e.g. *Yawn*, *Leaning in*).
- Future plans include multilingual annotators and culturally grounded expansions.

C Annotator Details and Guidelines

C.1 Annotator Selection

Experts We recruited 4 Ph.D. candidates in clinical psychology who routinely interpret nonverbal behaviour as part of their training and research. All expert annotators are fluent in English. To ensure fair compensation, we set a minimum rate of \$15 per hour.

Non-experts We additionally recruited 5 graduate students outside clinical psychology who demonstrated English proficiency sufficient for the task. They were compensated at the same minimum rate of \$15 per hour.

C.2 Procedure

To balance cognitive load with annotation quality, we adopted a subsampling strategy. Each annotator labelled an identical set of 50 items, enabling us to compute inter-annotator agreement while keeping the session manageable.

C.3 Interface

Annotations were collected with Label Studio⁵ using the interface shown in Figure 6.

D More Analysis

D.1 Categorical Performance Difference

Methods We group the 407 cues by anatomical region (Face, Arms, Legs, etc.), calculate each model's mean accuracy per region, and visualize the results in Figure 7.

Results Figure 7 shows:

- **Facial cues do not yield the highest accuracy**, contrary to common assumptions.
- **Arms, hips, and hands/fingers achieve relatively higher accuracy**, suggesting clearer mappings to descriptors.

D.2 Appearing Human Size

Methods We measure each clip's average human bounding-box area and compute the Pearson correlation with model accuracy.

Results

- The mean correlation coefficient is -0.005 ± 0.065 , indicating **virtually no relationship** between on-screen size and accuracy.

⁵<https://labelstud.io/>

Category	Mind-state labels
BELIEFS	confidence, self-assurance, trust, doubt, skepticism, suspicious, disbelief, certainty, confidence in telling the truth, belief in one’s statement, negative or worrisome thoughts
INTENTIONS	emphasis, accusing, desire to appear polite and agreeable, desire to appear more attractive, desire to drive home a point, trying to attract a potential mate, directing attention, open to response, actively participating, gesture to confide, intent, accusation or emphasis, joking gesture, stop-sign (blocking), signalling closeness, asking consent
PERCEPTS	attentive, attention, observing, focus, engagement, passive observation, distracted, disinterest, curiosity, showing focused attention, glare, looking away, openness, withdrawal
DESIRES	seeking comfort or reassurance, desire for self-comfort, desire for closeness and bonding, seeking understanding, desire to emphasize, desire to appear attractive, trying to block out pain, wanting privacy, wanting relief, yearning/intense wanting (energy)
KNOWLEDGE	uncertainty, genuine uncertainty (‘I really don’t know’), confusion, contemplation, thoughtfulness, reflection, consideration, awareness, evaluation / judging, realization, inquisitiveness
EMOTIONS	stress, anxiety, fear, panic, anger, annoyance, irritation, happiness, joy, sadness, calm, relaxation, affection, warmth, excitement, enthusiasm, nervousness, frustration, comfort, disgust, aversion, contempt, surprise, shock, embarrassment, humility, fatigue, tiredness

Table 5: Representative ‘explanation’ labels onto six broad cognitive–affective categories used in Theory-of-Mind literature (Ma et al., 2023). 🧠MOTION2MIND covers wide range of human cognition.

- **Bounding-box size has little influence** on model performance.

D.3 Frame Numbers

Methods We sample up to 32 frames at equal intervals (varying limits of 32, 16, 8, 4 frames) and measure model accuracy on Detection and Explanation tasks.

Results In Figure 8,

- **Accuracy declines as the number of frames decreases**, due to loss of motion clarity.
- **Explanation drops less sharply than Detection**, since it leverages script context.

E Experimental Setup

E.1 Hardware & Inference

- Up to 4× NVIDIA GeForce RTX 3090 GPUs for the 32B vision–language model
- 1× NVIDIA GeForce RTX 3090 GPU for all other models (8B, 7B, 3B, 2B, 1B)
- Paged attention via VLLM library (Kwon et al., 2023)
- Inference time: under 2 hours per task type

E.2 Hyperparameters

- Temperature: 0 (for deterministic outputs)
- Random seed: 0

- Maximum output tokens: varied across tasks and models

- Top- p sampling: 0.001

- Repetition penalty: 1.05

E.3 List of LLMs Used in Paper

The models we utilized in this paper are as follows:

- GPT-o1⁶
- GPT-4o / GPT-4o-mini⁷
- Gemini-1.5-Flash⁸
- Qwen2.5-VL-32B/7B/3B-Instruct⁹
- InternVL3-8B/2B/1B¹⁰

F Option Generation Algorithm

In §3 and §5, we formulate the benchmark as a multiple-choice reasoning task. Distractor options are constructed by sampling psychologically plausible but contextually inconsistent explanations from the same cue category.

Specifically, given a target cue and its correct explanation, we retrieve semantically distant explanations associated with the same nonverbal cue using GPT-text embedding cosine similarity. This

⁶GPT-o1

⁷GPT-4o; GPT-4o-mini

⁸Gemini-1.5-Flash

⁹Qwen2.5-VL-32B, 7B, 3B-Instruct

¹⁰InternVL3-8B, 2B, 1B

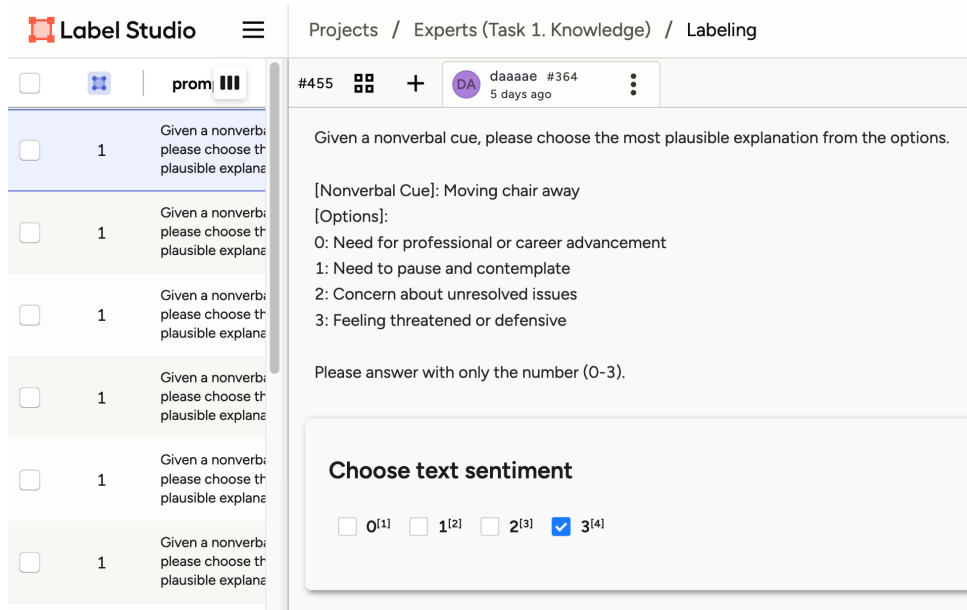


Figure 6: Example of the labeling interface.

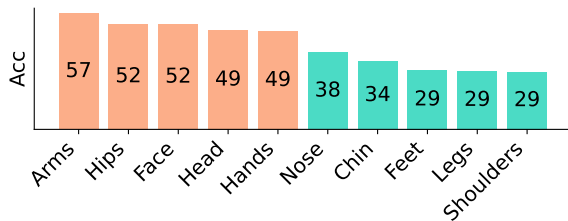


Figure 7: 5 most accurate (Orange) and inaccurate (Green) body parts. Models are less likely to choose ‘invalid’ responses when similar NVC is added to the dialogue (x: NVC numbers, y: Answer as invalid) for both validity and explanation tasks.

prevents trivial elimination based solely on cue mismatch while preserving semantic plausibility. This construction strategy increases reasoning difficulty by requiring contextual discrimination rather than simple cue recognition.

G Prompts

In Table 6 and Table 7, we specify the prompts we use for §4 and §5.

H Use of AI Assistants

AI assistants were used for coding support and minor grammatical correction. All experimental design, annotation protocols, analyses, and scientific claims were independently developed and verified by the authors.

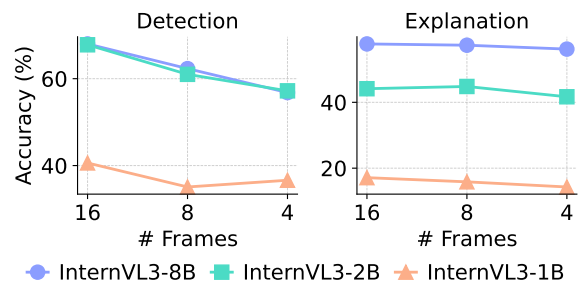


Figure 8: Accuracy versus maximum input frames.

Variable: body part, Frames {Frames} Please explain the nonverbal cues in the video of the given body part in the most detail. – If multiple people appear, explain each person’s cues separately. – Do <u>not</u> mention cues unrelated to the specified body part. [Body part]: {body part}
Variables: script + caption Given the caption about the short video clip and script, please parse the appearing nonverbal cues into JSON format. Do <u>not</u> annotate vocal cues. FORMAT: <pre>[{ "cue_id": "0", "cue_sign": "...", # concise description "body_part": "...", # head, face, neck, arms ... "cue_agent": "...", # who performed the action "mind_state": "...", # psychological meaning or "none" "detail": "...", # extra detail }, ...]</pre> [Script with Caption] {script + caption} [Appearing action]

Table 6: Captions used in §4. Prompt used to get nonvnerbal cue captions in the video and reconstruct the data into json format.

Variables: script, agent, options Given the following script and a <u>video clip</u>, please select the most plausible nonverbal action (behaviour by {agent}) in the blank. The MARKED SCENE is bounded by ***** SCENE START ***** and ***** SCENE END *****. The previous <u>chunk</u> of the scene is included for context. [Script] {script} Choose from the following options (answer only the option number without any other text): {options}
Variables: script, options Given the following script of a short video clip, please explain the nonverbal action in the blank. Focus on the cue between the scene start and end marks. [Script] {script} Choose from the following options (answer only the option number without any other text): {options}
Variables: script, options Given the following script of a short video clip, please predict the next utterance in the blank. Focus on the cue between the scene start and end marks. [Script] {script} Choose from the following options (answer only the option number without any other text): {options}
Variables: agent, options Given the following <u>video</u>, please detect what nonverbal cue (behaviour by {agent}) is present. Choose from the following options (answer only the option number without any other text): {options}
Variables: cue, agent, options Given the following <u>video</u>, please detect whether the specified nonverbal cue appears. Nonverbal cue: {cue} by {agent} Choose from the following options (answer only the option number without any other text): 1. appears 2. does not appear

Table 7: Prompt templates for the five task types used in our benchmark, ordered left-to-right: **cue**, **explanation**, **next_prediction**, **detection**, and **detection_binary**. Curly-braced tokens ({ . . . }) are filled at runtime.