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

Anonymous ACL submission

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 😱MOTION2MIND, a framework for evaluating the ToM capabilities of machines in interpreting NVCs. Leveraging an expert-curated body-language reference as a proxy knowledge base, we build 😱MOTION2MIND, a carefully curated video dataset with fine-grained nonverbal cue annotations paired with manually verified psychological interpretations. It encompasses 222 types of nonverbal cues and 397 mind states. Our evaluation reveals that current AI systems struggle significantly 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. We make our data and public.

1 Introduction

Understanding **others' mental states through visual cues** is fundamental to human social interaction and intelligence (Fernandez-Duque and Baird, 2005; 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.

Large Language Models (LLMs) have made remarkable progress in processing text-based interactions (Park et al., 2023), yet their capability to understand subtle mental states expressed through nonverbal communication remains largely unverified. Existing Theory of Mind (ToM) benchmarks (Le et al., 2019; Weber et al., 2021; Jin et al., 2024a) have advanced, but they primarily focus on false-belief tasks (Wimmer and Perner, 1983) - testing an agent's ability to reason about asymmetric **information** between characters. However, there is a growing body of work that calls for a broader spectrum of mental state inference in ToM tasks (Ma et al., 2023; Wang et al., 2025).

Another attempt to measure NVC understanding through video datasets (Luo et al., 2020; Chen et al., 2023; Liu et al., 2021a; Huang et al., 2021) has encountered two significant methodological limitations. First, they employ an oversimplified scoring system focused on emotions (e.g., rating valence/arousal on a 1-7 scale), which fails to capture the broad range of mental states. Secondly, most of these datasets span from several minutes to several hours, during which numerous NVCs appear, but individual annotations for each NVC are not provided.

To address these challenges, we introduce 😱MOTION2MIND, a comprehensive framework to evaluate mind interpretation capabilities using NVC as important information. Our framework is seeded by an expert-curated body-language reference that enumerates 407 frequently discussed cues and their plausible psychological interpretations. We use this reference as a structured prior against which we can measure how well models align with human-documented associations when applied to realistic contexts drawn from sitcoms, reality footage, and film. Our data is validated by a high accuracy of human annotators demonstrating its plausibility and clarity. While the current state-of-the-art model GPT-4o (OpenAI et al., 2024a) correctly guesses complex false belief tasks (Kosinski, 2024), it fails to align with even this day-to-day NVC knowledge in realistic contexts.

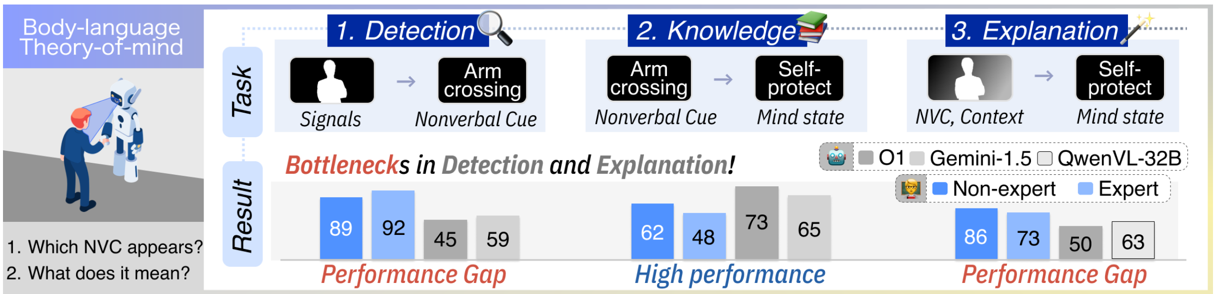

Figure 1: We disentangle concept of **nonverbal cue understanding** into three distinct components: (1) **Detection**, identifying and labeling various naturalistic movements; (2) **Knowledge**, the general understanding of the psychological meanings associated with specific cues; and (3) **Explanation**, contextual reasoning to infer the psychological state behind observed cues. Our test set, developed based on Joe Navarro's work, reveals that while LLMs perform comparably to humans in Knowledge, they exhibit a substantial gap in the Explanation and Detection phase.

Our key contributions are:

1. **A Three-Stage Framework for Nonverbal Theory of Mind.** As shown in Figure 1, we propose a structured framework for understanding nonverbal communication with three distinct components: *Detection*, *Knowledge*, and *Explanation*.

2. **MOTION2MIND: A Realistic, Multimodal Benchmark with Contextually Invalid Cues.** We operationalize the reference dictionary inside contextual video clips and additionally include 'invalid' cues (salient but carrying no dictionary-supported meaning) to test over-interpretation.

3. **Comprehensive Evaluation of Model Competence in Nonverbal Mind Inference.** We assess five task types to quantify how closely models reproduce documented cue and meaning, comparing against experts and non-experts.

In §2, we introduce key components for theorizing nonverbal cue (NVC) communication. §3 evaluates basic knowledge of the NVCs without contexts. §4 introduces our MOTION2MIND framework, and §5 presents empirical analyzes of current models.

2 Components in Understanding Nonverbal Theory of Mind

Many psychological studies divide the mentalization process into successive stages (Fonagy, 2011; Heider, 2013). 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

Detection converts raw multimodal signals into discrete nonverbal cue recognition. Accurate detection is a prerequisite for downstream inference. Key challenges include handling inter- and intra-subject variability and mitigating noise (*e.g.*camera angle, background audio).

2.2 Knowledge

The knowledge component maps each detected cue to a set of 'plausible' psychological meanings. Considering the nature of nonverbal cues, where a single cue can convey multiple meanings, psychological studies use patterns from various contexts. We reference an expert-curated body-language dictionary (Navarro, 2018) as a proxy knowledge base. It lists 400+ cues and multiple plausible interpretations. More analysis about the reference is in Appendix B.

2.3 Explanation

Explanation takes the candidate interpretations from the knowledge component and combines them with contextual information to yield a final mental-state hypothesis (e.g. 'surprised,' 'engaged'). 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 interpretation (emotion, attitude, or intention).

3 Knowledge: Body-language understanding Without Context

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

	Cue → **Explanation**	Explanation → **Cue**
Prompt	Given a nonverbal cue, please choose the most plausible explanation from the options. *'Arm crossing'*	Given the explanation of a nonverbal cue, please provide a plausible nonverbal cue from the options. *'Feeling insecure or threatened'*
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: Example of prompts in §3. We implement two-sided tasks: Cue to Explanation and Explanation to Cue.

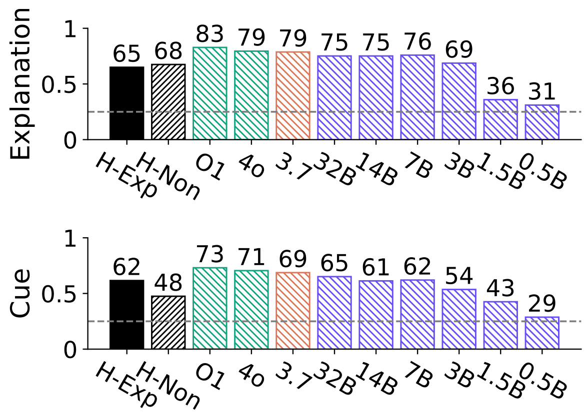

Figure 2: NVC knowledge scores of intelligent LLMs — GPT (green), Claude (orange), Qwen2.5-Instruct (purple) — tested on the NVC dictionary. LLMs manifest structurized knowledge even than psychological experts.

with the structured NVC dictionary by human experts (Navarro, 2018).

3.1 Methodology

Test Set Navarro (2018) covers 407 NVCs and their possible (multiple) psychological meanings. 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.

Given the multi-answer nature of NVC interaction, we simplify the task into Multi-choice QA questionnaires for clear evaluation. To construct meaningful but clear distractors, we use cosine similarity between semantic embeddings[1] and select options whose explanations are semantically distant from all explanation units associated with the correct answer (See Appendix F).

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 documented knowledge. This indicates that they possess a high level of structured knowledge about nonverbal cues. Human experts tend to struggle more, likely due to the absence of contextual information that typically aids interpretation.

Large models do better in both tasks. Our results show a clear scaling effect across models of different sizes. Larger models, such as GPT-o1 and Qwen2.5-32B, consistently outperform smaller ones in both understanding and generation tasks. This scaling trend indicates that larger models better align with the reference associations, suggesting the dictionary is a sufficiently structured prior for comparative evaluation.

Understanding > generation. Across all models, selecting the correct explanation for a given cue is generally easier than generating a cue based on a described mental state. Human participants also perform better in the understanding task, but the

[1]We use OpenAI's 'text-embedding-3-small' for computing semantic embeddings.

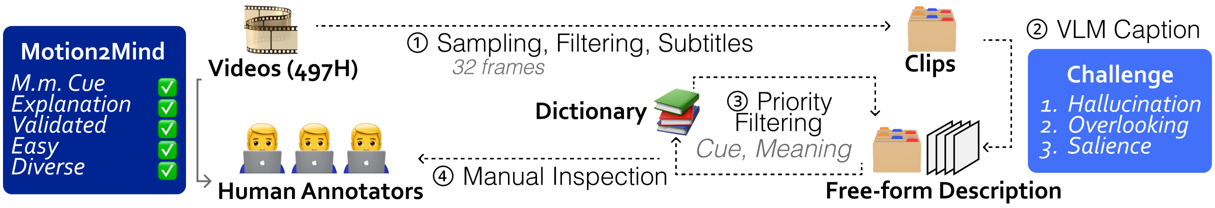

Figure 3: We build MOTION2MIND, a dataset annotated with fine-grained multimodal (m.m.) cues. To construct the dataset, we collect 497 hours of video from YouTube (sitcoms, movies, reality shows), sample short clips (32 frames), and generate initial captions using Qwen2.5-32B-VL-Instruct. These captions are filtered using a body-language dictionary to prioritize clips with interpretable cues and meanings. Human annotators then manually inspect the clips and refine the explanations based on contextual grounding, ensuring that each cue is paired with its most accurate and salient psychological meaning within the scene.

difference between understanding and generation is less pronounced compared to LLMs.

4 MOTION2MIND

We present 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 To address the limitations of prior NVC datasets (see Table 2), which often suffer from small scale or restricted annotation types, we source diverse videos from six high-subscriber YouTube channels spanning film, television, and reality genres. Using the YT-DLP framework (yt-dlp contributors, 2025), 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. To ensure fair coverage across different video types and lengths, we extract clips in proportion to video length and cap the number from each video at 40 (approximately half the mode of video lengths). 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 (Jocher et al., 2023) to detect human presence and track consistency across frames.

Subtitles We extract spoken dialogue using Whisper-large-v3 (Radford et al., 2022), and align utterances to video timestamps. Speaker segmentation is performed using NVIDIA NeMo (Kuchaiev et al., 2019), allowing us to associate vocal cues with specific individuals in each clip.

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 appearing cues or overlooking appearing cues. (2) *Misalignment with Human Salience*: The described cue is present but not the most overt for the human.

Solution 1: Body Part Detection and Prompt Conditioning We apply MediaPipe (Lugaresi et al., 2019) 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 and focused prompting to produce more specific descriptions.

Solution 2: Character-Specific Captioning We then generate visual descriptions for each individual in the clip. For each detected character and their visible body parts, we prompt the VLM to describe their behavior. Through prompt engineering, we find that simple instructions yield more accurate

Dataset	Items	Mods	# Mind	Cue.	Invalid	Vocal.	Source
MOTION2MIND	1,022	V +A +T	397	✓	✓	✓	Movie, Sitcom, Reality
SOCIAL GENOME (Mathur et al., 2025)	272	V +A +T	—	✓	✗	✓	YouTube
MMToM-QA (Jin et al., 2024b)	7.5k	V +A +T	Unk (B, D, I)	✓	✗	✗	Simulation
Aff–Wild2 (Kollias and Zafeiriou, 2019)	548	V +A	8 (E)	✗	✗	✓	YouTube
VEATIC (Ren et al., 2023)	124	V +A	Cont. (E)	✗	✗	✓	Mixed clips
MovieGraphs (Vicol et al., 2018)	7.6k	V +T	9 (R)	✓	✗	✗	Movies
Social–IQ (Li et al., 2025)	1.2k	V +T	QA	✓	✗	✗	YouTube
iMiGUE (Liu et al., 2021b)	359	V	3 (E)	✗	✗	✗	Tennis press
BoLD / ARBEE (Luo et al., 2019)	9.8k	V	26 (E)	✗	✗	✗	Movies
BoME (Wu et al., 2023)	1.6k	V	4 (E)	✗	✗	✗	AVA-derived

Table 2: We introduce MOTION2MIND, the first multimodal dataset with fine-grained motion annotations and validated psychological explanations. V = vision, A = audio, T = text. Cue. denotes specification of behavior in the visual modality. B, D, I, E, R stand for Belief, Desire, Intention, Emotion, and Relationship, respectively. Cont. = continuous variable; Vocal. = annotation of vocal nonverbal cue.

and informative results than complex task-specific prompts.

Solution 3: 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). Our dictionary is comprehensive that defines 407 validated nonverbal cues, and this step eliminates subjective or overly creative outputs and narrows the candidate set for human review.

Solution 4: 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 We measure words per minute (WPM) within each segment, dynamically applying the mean and standard deviation for speaker. We label `[Fast]` when normalized WPM exceeds 1.5.

Pitch We estimate pitch for each utterance using Parselmouth (Boersma and Weenink, 2021). Segments shorter than 120 ms are excluded for reliable estimation. Similarly with Speaking rate, we annotate `[HIGH_PITCH]` when normalized pitch surpasses 1.25.

Long Pause Silent periods are detected using WebRTC 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 a predefined body language dictionary containing 407 cue types and 2,050 possible psychological explanations. This ensures that all labels are grounded in established psychological literature. During manual inspection, we find that most explanations are grounded by the dictionary, showing its broad coverage.

Solution 2: Human-Guided Labeling and Invalid Cases Each cue is reviewed by human annotators using the dictionary as reference. Annotators select the most contextually appropriate explanation. We also incorporate ‘Invalid’ if the cue is apparent but not directly pointing any psychological state. To reduce *Subjectivity*, all annotations are cross-checked by a second annotator.

Model	Open	Input	ToM Method	Detection		Cue	Explanation			Prediction
				MCQ	Binary	Accuracy	Total	Valid	Invalid	MCQ
Expert	–	–	–	–	89.0	–	**81.3**	**76.3**	**86.3**	90.0
Non-expert	–	–	–	–	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)*	✗	65.0	69.3	47.7	59.6	65.5	30.0	83.2
Qwen 2.5-7B	✓	*V, T, (A)*	✗	67.6	32.3	46.8	59.5	65.1	29.6	49.5
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. (2023)	-	-	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. We evaluate each model across five tasks: (1) *Detection* (MCQ): Identify the correct nonverbal cue of video clip. (2) *Detection* (Binary): Determine whether a given cue appears in the clip. (3) *Cue*: Choose the most appropriate nonverbal cue that would occur in context. (4) *Explanation*: Infer the likely mind state of thelo given cue. (5) *Prediction*: Anticipate thioke next line of dialogue following a cue. VLMs consistently underperform humans across tasks. The random baseline is 25% for all multiple-choice tasks except for Detection - Binary. Input: V = visual (frames), T = text, A = audio features.

5 Test VLMs

We test current VLMs' performance on our MOTION2MIND benchmark. Specifically, we test GPT o1 and 4o, Qwen2.5-VL (Wang et al., 2024) 32B to 3B, and InternVL (Chen et al., 2024c) 8B to 2B. For a clear evaluation, we formulate this task as a multiple-choice question (MCQ) similar to §3, and the answer positions are randomized between four to eliminate position bias.

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.

5.2 Task Definition

Detection The goal is to identify which nonverbal cue appears in the given visual input. We design two formats for robustness: (1) MCQ, where the model selects the correct cue from multiple distractors (cues), and (2) Binary, where the frames and a candidate cue is provided, and the model chooses between 'Appears.' and 'Does not appear'.

Cue (Generation) The task is to infer the most plausible nonverbal cue in the blank. Since script-only lacks contextual information, we also provide the preceding 4-second video chunk to supply relevant visual context while avoiding spoilers.

Explanation 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 or emotional state. This task evaluates the model's ability to interpret the meaning of observed behavior.

Prediction (Next Utterance) This task provides both the visual clip and its surrounding script, with a blank for the next line of dialogue following the nonverbal cue. The model must choose the most plausible next utterance, serving 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 §5.4.

Theory of Mind Modules Yield Limited Gains While models with ToM modules (Wilf et al., 2023; Sclar et al., 2023; Jin et al., 2024a; Zhang et al., 2025) 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 Over- vs. Under-interpretation

In *Explanation* task, we categorize the combination of model answer type and ground-truth typology in Table 4.

Type	Ground-truth	Model answer
TP	Valid	Same Valid
FN	Valid	Invalid
EP	Valid	Different Valid
TN	Invalid	Invalid
FP	Invalid	A valid

Table 4: Ground-truth is labeled by human annotators, and 'valid' means that the NVC shows some distinct psychological meaning in the context (*e.g.*Stressed). We define False Negative (FN) and False Positive (FP) as under-interpretation and over-interpretation.

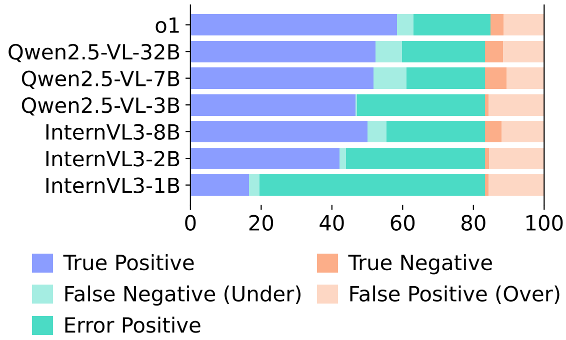

Figure 4: Stacked bar plots of *Explanation* task answers. Small models shows low precision (over-interpret) compared to larger models.

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.

5.5 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, such as just sitting forward alone is connected with 'intention to show empathy'.

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; Amirizaniani et al., 2024), some papers encompassing visual cues as input (Jin et al., 2024a; Chen et al., 2024a; Zhang et al., 2024; van Groenestijn, 2024; Etesam et al., 2023; Ma et al., 2023) evaluating models' ability to distinguish asymmetric information in templated stories. Recent efforts expand ToM assessments to broader mental 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;

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).

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., 2021a; Huang et al., 2021; Wicke, 2024; Zadeh et al., 2019; Lu et al., 2020; Chen et al., 2024b; Tapaswi et al., 2019). 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 NVCs.

Affective Computing & HRI Affective HRI aims to sense and react to human states from facial, bodily, and vocal cues (Picard, 1997; Spezialetti et al., 2020). Early work centered on real-time emotion or intent recognition for assistive robots (Rudovic et al., 2018; van der Pol et al., 2022). Recent studies embed explicit ToM: false-belief reasoning on humanoids (Zeng et al., 2020) and GPT-4V–based multimodal inference in AToM-Bot (Shu et al., 2024), advancing toward robots with functional Theory of Mind (Breazeal and Scassellati, 2002; Sturgeon et al., 2021).

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 significantly when faced with contextual ambiguity and nuanced social 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

Proxy Nature of the Dictionary Our annotations and *Knowledge* tasks are grounded in a single expert-curated body-language dictionary. We use it as a proxy knowledge base as the most diverse and extensive NVC reference we are aware of. As psychological definitions of nonverbal behaviors remain fluid and debated, especially for new gestures and micro-expressions identified, our fixed taxonomy may not capture them.

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.

References

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. arXiv preprint arXiv: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.

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/.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. 2024. Explicit modelling of theory of mind for belief prediction in nonverbal social interactions. arXiv preprint arXiv:2407.06762.

Cynthia Breazeal and Brian Scassellati. 2002. Using robots to study joint attention. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1650–1655.

Judee K. Burgoon, Laura K. Guerrero, and Kory Floyd. 2016. Nonverbal Communication, 2nd edition. Routledge.

Judee K. Burgoon and Valerie Manusov. 1994. The arm-crossed gesture: A nonverbal cue of attitude? Journal of Nonverbal Behavior, 18(4):261–278.

Dana R. Carney, Amy J.C. Cuddy, and Andy J. Yap. 2010. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. Psychological Science, 21(10):1363–1368.

Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. International Journal of Computer Vision, 131(6):1346–1366.

Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2024a. Through the theory of mind's eye: Reading minds with multimodal video large language models. arXiv preprint arXiv:2406.13763.

Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2024b. Through the theory of mind's eye: Reading minds with multimodal video large language models. arXiv preprint arXiv:2406.13763.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271.

Charles Darwin. 1872. The Expression of the Emotions in Man and Animals. John Murray.

John Davis and Brenda Smith. 2010. Forehead wrinkles and perceived age: A field study. Journal of Cosmetic Dermatology, 9(4):274–278.

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.

Paul Ekman. 1997. Facial expressions of emotion: New findings, new questions. Psychological Science, 3(1):34–38.

Paul Ekman. 2003. Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life. Times Books.

Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2023. Emotional theory of mind: Bridging fast visual processing with slow linguistic reasoning. arXiv preprint arXiv:2310.19995.

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.

Julius Fast. 1970. Body Language. Simon and Schuster.

Diego Fernandez-Duque and Jodie A Baird. 2005. Is there a 'social brain'? lessons from eye-gaze following, joint attention, and autism. Other minds: How humans bridge the divide between self and others, pages 75–90.

Peter Fonagy. 2011. The mentalization-focused approach to social development. In Mentalization, pages 3–56. Routledge.

Jonathan B Freeman and Nalini Ambady. 2011. A dynamic interactive theory of person construal. Psychological review, 118(2):247.

David Givens. 2016. Nonverbal Dictionary of Gestures, Signs and Body Language Cues. Center for Nonverbal Studies.

Fritz Heider. 2013. The psychology of interpersonal relations. Psychology Press.

Ursula Hess, Reginald B Adams Jr, and Robert E Kleck. 2010. The influence of intensity, gender, and sex of the encoder on judgments of dominance and affiliation from dynamic emotional expressions. Journal of Nonverbal Behavior, 34(4):259–269.

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 B Tenenbaum, and Tianmin Shu. 2024a. Mmtom-qa: Multimodal theory of mind question answering. arXiv preprint arXiv:2401.08743.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024b. Mmtom-qa: Multimodal theory of mind question answering. Preprint, arXiv:2401.08743.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics yolov8.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. arXiv preprint arXiv:2310.15421.

Chris L Kleinke. 1986. Gaze and eye contact: A research review. Psychological Bulletin, 100(1):78–100.

Mark L. Knapp and Judith A. Hall. 2007. Nonverbal Communication in Human Interaction, 7th edition. Wadsworth.

Dimitrios Kollias and Stefanos Zafeiriou. 2019. Aff-wild2: Extending the aff-wild database for affect recognition. Preprint, arXiv:1811.07770.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. Proceedings of the National Academy of Sciences, 121(45):e2405460121.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. `https://github.com/NVIDIA/NeMo`. ArXiv preprint arXiv:1909.09577.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.

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.

Hao Li, Hao Fei, Zechao Hu, Zhengwei Yang, and Zheng Wang. 2025. Vegas: Towards visually explainable and grounded artificial social intelligence. Preprint, arXiv:2504.02227.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021a. 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.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhaoz? 2021b. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. Preprint, arXiv:2107.00285.

Chih-Yuan Lu, Fangyu Huang, Chenyu Tan, Richard Wang, and Wonmin Byeongho Choi. 2020. Video-and-language event prediction (vlep). In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10654–10661.

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. arXiv preprint.

Yu Luo, Jianbo Ye, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. 2019. Arbee: Towards automated recognition of bodily expression of emotion in the wild. International Journal of Computer Vision, 128(1):1–25.

Yu Luo, Jianbo Ye, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. 2020. Arbee: Towards automated recognition of bodily expression of emotion in the wild. International journal of computer vision, 128:1–25.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. arXiv preprint arXiv:2310.19619.

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 AAAI Conference on Artificial Intelligence.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. arXiv preprint arXiv:2404.11023.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. arXiv preprint arXiv:2502.15109.

David Matsumoto and Hyisung C Hwang. 2008. Recognizing facial expressions of emotion. Psychological Science Agenda, 22(2):6–7.

Albert Mehrabian. 1972. Nonverbal communication. Transaction Publishers.

Desmond Morris. 1977. Manwatching: A Field Guide to Human Behavior. Harry N. Abrams.

Joe Navarro. 2018. The Dictionary of Body Language: A Field Guide to Human Behavior. Harper Collins.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai

Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. Preprint, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024b. Openai o1 system card. Preprint, arXiv:2412.16720.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pages 1–22.

Allan Pease and Barbara Pease. 2004. The Definitive Book of Body Language. Bantam.

Rosalind W. Picard. 1997. Affective Computing. MIT Press, Cambridge, MA.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. Preprint, arXiv:2212.04356.

Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X. Yu, and David Whitney. 2023. Veatic: Video-based emotion and affect tracking in context dataset. Preprint, arXiv:2309.06745.

Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. Science Robotics, 3(16):eaar6760. ArXiv:1802.01186.

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. arXiv preprint arXiv:2306.00924.

Tianmin Shu, Olivier Pietquin, Anu Radha, Michael Chu, Sanjeev Mohan, and Joshua B. Tenenbaum. 2024. Atom-bot: Affective theory of mind for empathetic human–robot interaction. In arXiv preprint arXiv:2406.08455v2.

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.

Stephanie Sturgeon, Andrew Palmer, Janelle Blankenburg, and David Feil-Seifer. 2021. Perception of social intelligence in robots performing false-belief tasks. Human–Robot Interaction, 10(1):45–60.

Makarand Tapaswi, Yuanjun Zhu, and Rainer Stiefelhagen. 2019. Moviegraphs: Towards understanding human-centric situations from videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1662–1671.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

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.

Ruth Tramposch and William Hart. 2021. Power postures versus expansive postures: Which posture displays dominance best? Personality and Individual Differences, 169:110066.

E. van der Pol, J. K. Karemaker, and B. van Arem. 2022. Vision-based intent prediction in social navigation scenarios. Robotics and Autonomous Systems, 147:103851.

AM van Groenestijn. 2024. Investigating theory of mind capabilities in multimodal large language models.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. Preprint, arXiv:1712.06761.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. Preprint, arXiv:2409.12191.

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.

Manuel Weber, David Kersting, Lale Umutlu, Michael Schäfers, Christoph Rischpler, Wolfgang P Fendler, Irène Buvat, Ken Herrmann, and Robert Seifert. 2021. Just another "clever hans"? neural networks and fdg pet-ct to predict the outcome of patients with breast cancer. European journal of nuclear medicine and molecular imaging, pages 1–10.

Philipp Wicke. 2024. Probing language models' gesture understanding for enhanced human-ai interaction. arXiv preprint arXiv:2401.17858.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. arXiv preprint arXiv:2311.10227.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1):103–128.

Chenyan Wu, Dolzodmaa Davaasuren, Tal Shafir, Rachelle Tsachor, and James Z. Wang. 2023. Bodily expressed emotion understanding through integrating laban movement analysis. Preprint, arXiv:2304.02187.

Ananya Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the gesture: Evaluating ai sensitivity to culturally offensive non-verbal gestures. arXiv preprint arXiv:2502.17710.

yt-dlp contributors. 2025. yt-dlp: A feature-rich command-line audio/video downloader. https://github.com/yt-dlp/yt-dlp. Version 2025.04.30 (commit b77e5a5), accessed 2025-05-20.

Amirhossein Zadeh, Xi Chen, Soujanya Poria, 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 (CVPR), pages 8808–8818.

Yi Zeng, Yuxuan Zhao, Tielin Zhang, Dongcheng Zhao, Feifei Zhao, and Enmeng Lu. 2020. A brain-inspired model of theory of mind. Frontiers in Neurorobotics, 14:60.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Yongshan Chen, Landi Gao, Dakuo Wang, Weinan Zhang, Xinbing Wang, and Ying Wen. 2024. Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. arXiv preprint arXiv:2409.08811.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. arXiv preprint arXiv:2502.15676.

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**: 35% pass rate; 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 Coverage Against Literature

Great comprehensiveness To verify the comprehensiveness of our dictionary, we compare our cue inventory with five foundational NVC sources. Table 6 below summarizes which anatomical categories appear across sources and highlights representative cues.

Why Use a Single Dictionary?

While numerous works contribute valuable insights, we select this single dictionary as our primary annotation backbone for the following reasons:

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.

Anatomical Region	Cues	Sources
Eyes	Pupil dilation, gaze holding/averting	Darwin (1872), Ekman (2003), Kleinke (1986), Pease and Pease (2004), Knapp and Hall (2007)
Nose	Nostril flare, nose wrinkle	Ekman (1997), Pease and Pease (2004)
Mouth & Lips	Lip compression, lip purse, one-sided raise (AU14), jaw clench	Ekman (2003), Matsumoto and Hwang (2008), Hess et al. (2010), Fast (1970), Navarro (2018)
Cheeks & Jaw	Cheek sucking, cheek tension	Navarro (2018), Burgoon et al. (2016)
Eyebrows & Forehead	Inner raise (AU1+2), brow-lowering (AU4), forehead tension	Ekman (2003), Knapp and Hall (2007), Davis and Smith (2010), Fast (1970)
Head	Nods, shakes, head tilt, head turns	Pease and Pease (2004), Knapp and Hall (2007), Fast (1970)
Arms & Hands	Arms crossed, hands-akimbo, pointing, self-touch	Morris (1977), Burgoon and Manusov (1994), Carney et al. (2010), Pease and Pease (2004), Knapp and Hall (2007)
Shoulders & Torso	Shoulder shrug, slump, chest expansion	Darwin (1872), Mehrabian (1972), Tramposch and Hart (2021), Knapp and Hall (2007)
Pelvis & Hips	Pelvic retreat, pelvis forward	Givens (2016), Pease and Pease (2004)
Legs & Feet	Leg uncrossing, foot tapping, weight shifts	Morris (1977), Knapp and Hall (2007), Pease and Pease (2004)

Table 6: Overlap between our dictionary and foundational NVC literature, grouped by anatomical region.

- **Breadth**: Includes over 400 cues spanning full-body nonverbal expression.
- **Psychological Grounding**: Using a unified dictionary avoids inconsistency, semantic drift, and label redundancy.
- **Operational Clarity**: Offers standardized labels and illustrated guidelines.

B.3 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 (Yerukola et al., 2025).

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

	ToM Method	Cue	Explanation			Pred.
		Acc	Total	Val.	Inv.	Acc
32B	✗	47.7	59.6	65.5	30.0	**83.2**
	Wilf et al. (2023)	**59.2**	**61.8**	**65.6**	**40.0**	67.0
	Sclar et al. (2023)	48.6	45.4	49.6	21.3	53.0
	Jin et al. (2024a)	30.2	27.4	27.5	26.8	28.1
	Zhang et al. (2025)	40.1	50.7	56.1	24.6	37.7
7B	✗	46.8	59.5	65.1	29.6	49.5
	Wilf et al. (2023)	51.3	**63.8**	**68.0**	**40.0**	59.4
	Sclar et al. (2023)	**58.3**	51.4	56.0	25.3	**64.9**
	Jin et al. (2024a)	54.9	43.1	50.7	3.3	58.1
	Zhang et al. (2025)	43.0	47.7	54.7	11.4	63.2

Table 7: Performance of Qwen 2.5 models on Motion2Mind with four Theory of Mind methods (Wilf et al., 2023; Sclar et al., 2023; Jin et al., 2024a; Zhang et al., 2025). Pred. = Prediction, Val. = Valid, Inv. = Invalid.

	o1	Qw-32B	Qw-7B	Int-8B
By Situation Sentiment				
Negative	68.3	65.2	66.3	57.4
Neutral	75.0	63.7	67.2	64.1
Positive	75.5	65.9	71.5	66.1
By Social Situation				
Dialogue	77.8	78.6	94.7	70.0
Group Discussion	74.9	66.6	68.8	59.4
Intimate Conversation	74.8	62.4	66.2	65.6
Monologue	73.8	64.1	67.6	63.3
Multi-person Scene	78.9	71.1	85.7	85.2

Table 8: Model accuracies by situation sentiment (top block) and by social scenario (bottom block).

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 Studio[2] using the interface shown in Figure 6.

D More Analysis

D.1 ToM Methods

D.2 Cue Types and Social Scenario Types

Methods We classify each clip's sentiment as `Negative`, `Neutral`, or `Positive` by matching its script, mind-state label, and NVC explanation against predefined keyword sets. Simultaneously, we extract social-context features (number of speakers, presence of "?", "!", and scene markers) to assign each clip to one of: Dialogue, Monologue, Group Discussion, Intimate Conversation, or Multi-person Scene. For each model and category, we compute accuracy as the fraction of correctly answered MCQs.

Results Table 8 shows:

- **Neutral and Positive contexts yield higher accuracies than Negative ones**, and **o1 remains most robust to sentiment shifts**.
- **Dialogue and moderate-sized groups (2–4 speakers) achieve peak performance**.
- Accuracy drops in Monologues (1 speaker) and very large groups (> 5 speakers).

[2] https://labelstud.io/

D.3 Knowledge: Validity-Binary Task

Methods We sample each dictionary cue–explanation pair as a positive example and create a negative example by choosing a semantically distant explanation for the same cue. Models predict True/False and we measure accuracy, precision, and recall against these labels.

Model	Acc.	Prec.	Recall
Qwen2.5-32B-Instruct	0.886	0.964	0.834
Qwen2.5-14B-Instruct	**0.911**	0.923	0.902
Qwen2.5-7B-Instruct	0.875	0.927	0.839
Qwen2.5-3B-Instruct	0.894	0.856	**0.926**
Qwen2.5-1.5B-Instruct	0.884	**0.998**	0.814
Qwen2.5-0.5B-Instruct	0.565	0.966	0.536

Table 9: Validity-binary task: whether the cue–explanation pair is valid (random baseline = 0.5).

Results Table 9 indicates:

- **All models 1 B achieve >0.80 accuracy, precision, and recall**, demonstrating strong cue–meaning knowledge.
- **Performance drops sharply for the 0.5 B model**, highlighting the impact of model size on semantic understanding.

D.4 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.

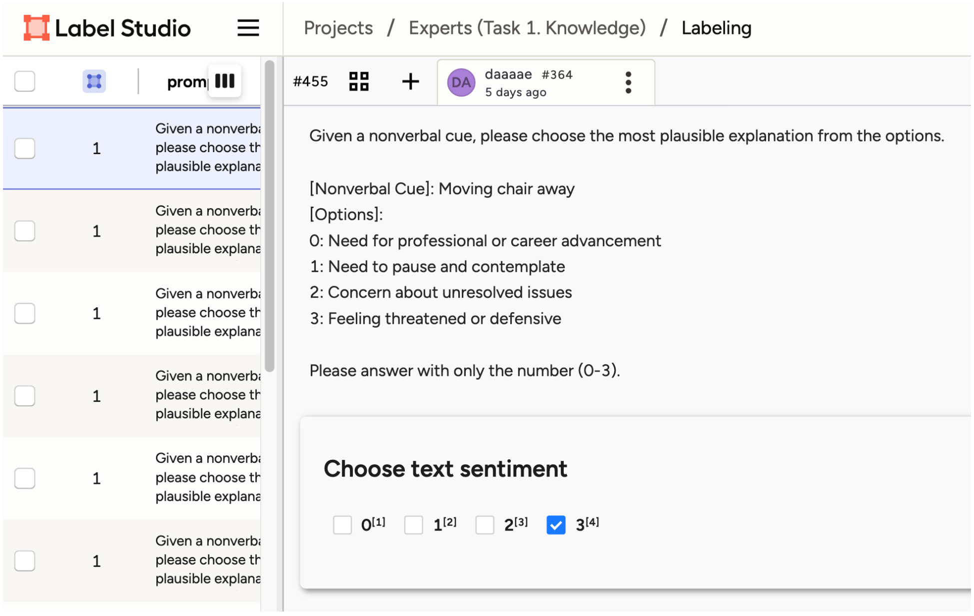

Figure 6: Example of the labeling interface.

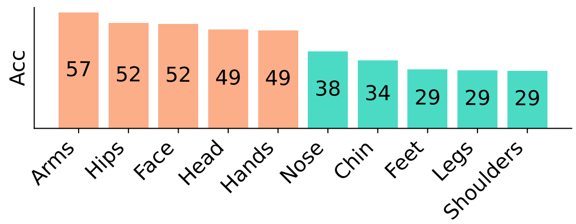

Figure 7: 5 most accurate (Orange) and inaccurate (Green) body parts. Models are less likely to choose 'invalid' responses when similar NVC is added to the dialogue (x: NVC numbers, y: Answer as invalid) for both validity and explanation tasks.

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.5 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.
- **Bounding-box size has little influence** on model performance.

D.6 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.

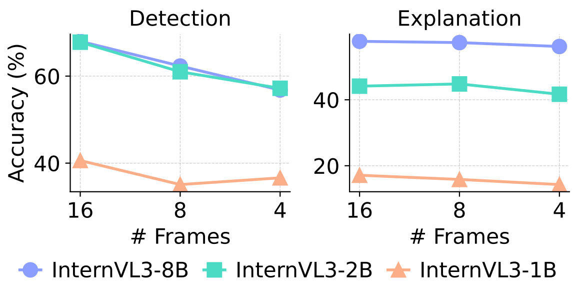

Figure 8: Accuracy versus maximum input frames.

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: variant
- 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-o1 (OpenAI et al., 2024b)
- GPT-4o (OpenAI et al., 2024a)
- GPT-4o-mini (OpenAI et al., 2024a)
- Gemini-1.5-Flash (Team et al., 2024)
- Qwen2.5-VL-32B-Instruct (Wang et al., 2024)
- Qwen2.5-VL-7B-Instruct (Wang et al., 2024)
- Qwen2.5-VL-3B-Instruct (Wang et al., 2024)
- InternVL3-8B (Chen et al., 2024c)
- InternVL3-2B (Chen et al., 2024c)
- InternVL3-1B (Chen et al., 2024c)

F Option Generation Algorithm

In §3 and §5, we utilze testset as multi-choice question format sourcing distractor options in the data pool. We use the semantic cosine distance, considering all the explanation pool described in dictionary given one nonverbal cue.

G Prompts

In Table 10 and Table 11, we specify the prompts we use for §4 and §5.

H Use of AI Assistants

We use AI assistants in coding and correcting grammatical errors.

Algorithm 1 GENDIVERSEOPTIONS

T: list of targets
I: list of items (each has *pivot*, *subcat*)
k: #options to pick (≈ 3)
dir $\in \{\text{far}, \text{close}\}$: choose dissimilar or similar distractors
$\tau_{\min}, \tau_{\max}$: cosine-similarity thresholds (optional) $\mathcal{R}$: MCQ records
Pre-compute embeddings
$C \leftarrow$ list of all *pivot* texts in I
$E \leftarrow$ ENCODE(C) *matrix $|I| \times d$
$t \in T$ $e^{\star} \leftarrow$ ENCODE($t.pivot$)
$\sigma \leftarrow$ cos_sim$(E, e^{\star})$ *$|I|$ scores
Candidate mask
mask $\leftarrow$ $\mathbf{true}^{|I|}$

if use subcategory **then** *mask* $\&= (I.subcat = t.subcat)$ exclude the target itself *mask* $\&= \big(C \neq t.pivot\big)$

if $\tau_{\min}$ given **then** *mask* $\&= (\sigma \geq \tau_{\min})$
if $\tau_{\max}$ given **then** *mask* $\&= (\sigma \leq \tau_{\max})$
$\mathcal{A} \leftarrow$ indices where *mask*$=$ true

if $|\mathcal{A}| < k$ **then***fallback $\mathcal{A} \leftarrow \{j \mid C[j] \neq t.pivot\}$
Greedy selection
$\mathcal{S} \leftarrow [\,]$

while $|\mathcal{S}| < k$ **do** dir $=$ far pick $j^{\star} = \arg\min_{j \in \mathcal{A}} \sigma[j]$ pick $j^{\star} = \arg\max_{j \in \mathcal{A}} \sigma[j]$
$\mathcal{S} \mathrel{+}= [\,I[j^{\star}]\,]; \quad \mathcal{A} \leftarrow \mathcal{A} \setminus \{j^{\star}\}$
Assemble MCQ entry
$\mathcal{R} \mathrel{+}= \big\langle t,\, [\,t\,] \cup \mathcal{S} \big\rangle$ **return** $\mathcal{R}$

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 not 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 not annotate vocal cues.

FORMAT:

```
[
    {
        "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
    },
    ...
]
```

[Script with Caption]
`{script + caption}`

[Appearing action]

Table 10: Captions used in §4. Prompt used to get novnerbal cue captions in the video and reconstruct the data into json format.

Variables: `script`, `agent`, `options`

Given the following script and a video clip, 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 chunk 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 video, 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 video, 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 11: 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.