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 comprehensive framework for evaluating the ToM capabilities of machines in interpreting NVCs. Starting from an FBI agent's validated profile handbook, we develop 👁MOTION2MIND, a carefully curated video dataset with fine-grained annotations of NVCs paired with 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, but also 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 (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) advance, 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 papers which call for a much broader spectrum of mental state inference in ToM task (Ma et al., 2023; Wang et al., 2025).

Another attempt to measure NVC understanding capability 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 fail to capture the broad range of mental states. Second, these annotations lack pinpointed behavioral annotation -for instance, they lack information which identifies which exact moment in a video sequence indicates that a subject is in 'happiness' or 'proud of themselves'.

To address these challenges, we introduce 👁MOTION2MIND, a comprehensive framework to evaluate mind state interpretation capabilities using NV as important information. Our framework starts from an expert-established psychological literature about NVCs, and we expand into 👁MOTION2MIND, grounded in realistic contexts from sitcom, reality, and movie. Our data is validated by a high score of human labelers showing its plausibleness and clarity. While the current state-of-the-art model GPT-4o (OpenAI et al., 2024a) correctly guesses complex false belief tasks, it fails to understand day-to-day NVC in real-world simulating contexts.

Our key contributions are:

1. **👁MOTION2MIND: A Comprehensive Video Benchmark for Nonverbal Cue Anal-**

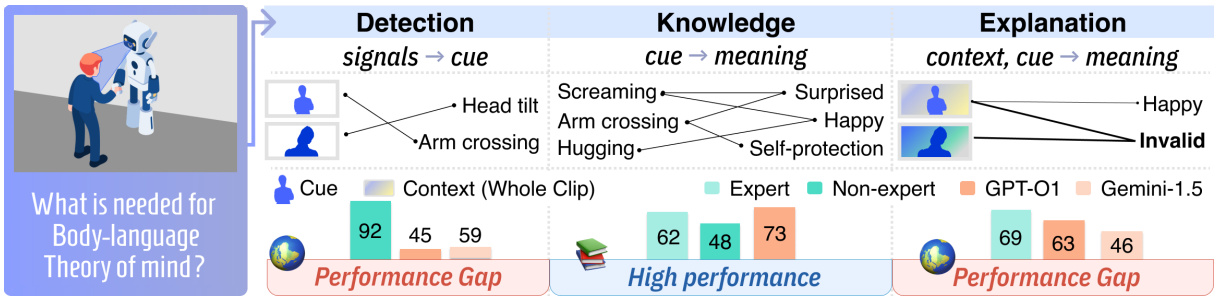

Figure 1: We disentangle concept of **nonverbal cue understanding** into three distinct components: (1) **Detection**, identifying and labeling various naturalistic movements; (2) **Knowledge**, the general understanding of the psychological meanings associated with specific cues; and (3) **Explanation**, 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.

ysis. We introduce a dataset comprising 1k annotated video clips with 222 unique nonverbal cues (e.g., neck stretching, high voice pitch) mapped to 397 psychological states.

2. **Contextual Analysis of NVC Interpretation in Vision-Language Models.** Through empirical analysis, we assess the extent to which state-of-the-art VLMs accurately identify and interpret nonverbal cues in varying social contexts, revealing a tendency toward over-interpretation.

3. **Bottleneck Identification in NVC Reasoning and Psychological Inference.** We pinpoint critical bottlenecks in current VLMs' interpretation of ambiguous cues, across three distinct components.

In §2, we introduce key components in theorizing Nonverbal cue (NVC) communications. §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 typically divide mentalization process into successive stages (Fonagy, 2011; Heider, 2013). To systemically 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. As shown in Figure 1, each nonverbal cue maps many-to-many to its meanings. There are some psychological studies establish logical foundations for interpreting nonverbal cues by using patterns from various contexts. We build our knowledge base on an expert-curated body-language dictionary authored by an experienced FBI agent (Navarro, 2018).

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). Unless noted, **ToM accuracy** refers to choosing the correct mind state among four options (§3.1).

3 Knowledge: Body-language understanding Without Context

We test state-of-the-art LLMs (GPT, Claude, Qwen2.5-Instruct) about the body language of

	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.

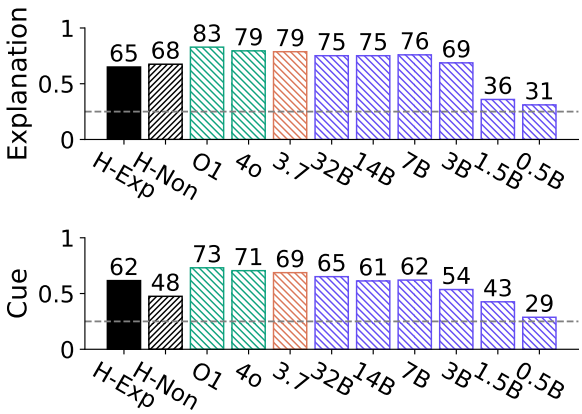

Figure 2: NVC knowledge scores of intelligent LLMs (GPT (green), Claude (orange), Qwen2.5-Instruct (purple)) tested on body language dictionary. LLMs manifests structurized knowledge even than psychological experts with clear scale effects.

others, using a validated body language dictionary (Navarro, 2018).

3.1 Methodology

Test Set We use the Body language dictionary (Navarro, 2018) as test set. This book covers 407 NVCs and their possible (multiple) psychological meanings. We structure the consolidated Explanation paragraph into n different semantic units (*e.g.*1. Stressed, 2. Threatened) using the 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. With cosine similarity of the semantic embeddings[1], we deliberately select distractor options with semantically distant *explanations* from the all the *explanation* units of correct answer. More details are in Appendix G.

Human Baselines Performance is measured against two human groups: Experts (psychologists with counseling certificates) and Non-Experts (general population). This dual baseline highlights gaps between LLMs and human understanding.

3.2 Results

Advanced Knowledge In Figure 2, all tested models significantly outperform even psychological experts, showing a strong ability in documented theoretical knowledge. They show scale effect, that larger models show better theoretical understanding of NVCs (*Explanation*: o1-83%, 32B-75%, 0.5B-31%). In our erroneous study, their erroneous answers are often plausible, not being entirely baseless.

Understanding vs. Generating Models get better scores at *understanding* cues (Explanation) than *generating* plausible cues (Cue) (83% vs 73% in O1). For instance, while 'Arm crossing' was correctly linked to 'Threatened', the reverse task yielded lower precision. This asymmetry mirrors human expertise, where decoding nonverbal signals is often easier than producing context-appropriate ones.

[1]Semantic embeddings in this paper use 'text-embedding-3-small'.

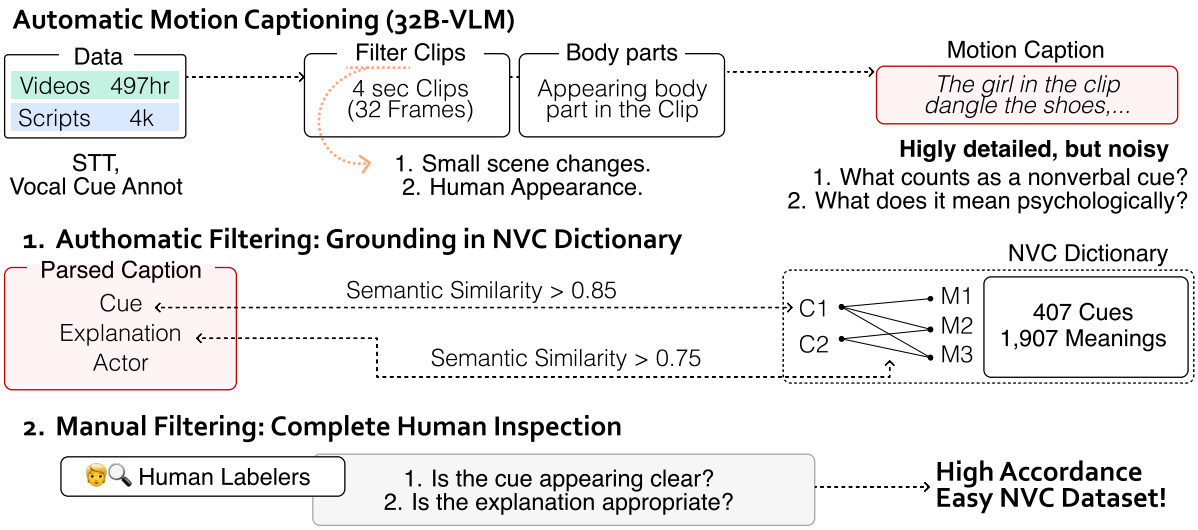

Figure 3: We build MOTION2MIND, a dataset with fine-grained nonverbal cue and validated psychologcial explanation. We source dataset from youtube (sitcom, movie, reality) and automatically process with video language model's captioning ability and dictionary's knowledge.

4 MOTION2MIND

We present MOTION2MIND, a carefully curated video dataset designed to assess psychological interpretation of nuanced body language within contexts. Our automated and scalable pipeline consists of distinct stages, systematically producing video clips (§4.1) annotated with nonverbal cues (§4.2) and corresponding psychological interpretations (§4.3).

4.1 Video Clips

Video Collection We collect video clips from YouTube channels spanning various genres, including sitcoms (Clipzone Sitcoms, 2025; The Office, 2025; Friends, 2025), movies (lionsgate, 2025; joblo, 2025), and reality shows (Keeping Up with the Kardashians, 2025), utilizing the yt-dlp (yt-dlp contributors, 2025) framework. These sources provide a wide array of social interactions and nonverbal behaviors, resulting in a total of 497.92 hours of videos divided into 4,730 clips.

Frame Extraction For every video, we randomly sample up to 40 clips with 4-second length. Each clip contains 32 frames captured at 8 fps, and this resolution is determined by our hands-on experience to sufficiently comprehend a whole appearance of nonverbal cue while maintaining efficiency. We employ Yolov8 (Jocher et al., 2023) to detect the number of people present in each frame and filter out (1) non-human content (e.g., cars, alarms) and (2) clips with frequent changes in the number of individuals, which indicates many scene transitions.

Subtitles Speech-to-text conversion is performed using Whisper-large-v3 (Radford et al., 2022), with speaker segmentation managed by Nvidia-Nemo (Kuchaiev et al., 2019). The resulting subtitles are then aligned with video timestamps.

4.2 Nonverbal Cue

Visual Cue Annotation As shown in Figure 3, we leverage Qwen2.5-32B-VL-Instruct to automatically generate descriptions of prominent nonverbal cues within each 4-second clip.

1. **Body Part Specification:** We use pose estimator model (Lugaresi et al., 2019) to identify appearing body parts (*e.g.*Face, Arms, Feet) within each clip. These components are then included in the text prompt to provide more granular annotations.

2. **Captioning:** Qwen-32B-VL-Instruct is prompted to generate separate descriptions for each detected person.

3. **Structurization:** GPT-4o-mini structures the generated free-form captions into a standardized JSON format, which includes the cue, actor, and the inferred mental state (if specified), and contextual details.

4. **Automatic Filtering:** We validate the cues using a predefined body language dictio-

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.

nary (Navarro, 2018), applying semantic and lexical matching with thresholds, respectively.

5. **Human Inspection:** The authors conduct a manual review to verify the appropriateness of detected cues, ensuring dataset reliability and content validity.

Vocal Cue: Automatic Pitch, Volume, Speech Speed Annotation Our vocal cue annotation pipeline identifies three primary vocal cues — speaking rate, pitch, and silence duration — by adapting baseline statistics per speaker.

1. **Speaking Rate:** The speaking rate is calculated as words per minute (WPM) within each segment. We update the baseline mean and standard deviation by the speaker. A segment is labeled as `[FAST]` if its WPM exceeds the baseline 1.5 times of mean by one standard deviation.

2. **Pitch:** Pitch estimation is conducted using Parselmouth (Boersma and Weenink, 2021), which applies Praat's pitch extraction algorithm. Segments shorter than 120 ms are excluded to prevent unreliable pitch estimation. If a segment's average pitch (F0) surpasses 1.25 times of the speaker's baseline mean by one standard deviation, it is labeled as `[HIGH_PITCH]`.

3. **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 Psychological Meaning (Explanation)

General Meaning As we do automatic filtering in §4.2, all the annotated nonverbal cue is mapped with a nonverbal cue name defined in dictionary, also with the multiple possible explanations. This explanation is conditioned with the nonverbal cue, not with the context.

Context-Dependent Meaning To ensure the selection of the most definitive subset, we adopt the following assumption and pseudo-labeling strategy, conducting manual annotation through all data items:

1. **Meaning Constraint:** We only acknowledge meanings present in the dictionary. While actions such as closing one's eyes can signify fatigue, stress, emotional response, or social etiquette (e.g., during a kissing moment), interpretations beyond the dictionary scope are excluded to minimize excessive subjectivity of annotations.

2. **LLM-Dictionary Alignment:** We prioritize labeling samples where mind state in §4.2-Structurization closely matches with the explanations in our dictionary. If the motion caption and the meaning are both have similarity with pain in dictionary, we consider it more compatible.

5 Test VLMs

Now we test current VLMs' performance with MOTION2MIND. We test current video language models (GPT-4o series, Qwen2.5-VL (Wang et al., 2024) series, and InternVL (Chen et al.,

Model	Open	Input	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
InternVL3-1B	✓	*V + T + (A)*	40.6	49.8	25.7	17.8	20.2	5.3	54.4

Table 3: Performance of VLMs on MOTION2MIND. VLMs generally perform worse than humans across Detection, Explanation, and Prediction tasks. Except Detection-Binary, the random baseline is 25.0% since we provide four options for each question.

2024c) series). For clear evaluation, we formulate this task as a multiple-choice question (MCQ), similar to §3, employing the same option sampling methods. We shuffle answer labels to remove label bias.

5.1 Task Definition

Detection (MCQ) Detection indentifies which nonverbal cue is present in a video clip. The inputs are raw multimodal signals, and the output is the specified cue.

Detection (Binary) To more clearly assess the detection ability, we ask model to determine whether a given cue appears in a short video using two options: "1. Appears" or "2. Does not appear.". The answer is always 'yes,'.

Cue Generation Cue generation identifies which nonverbal action should occur in a marked video segment. This task is akin to generating an appropriate nonverbal cue for the given scene. To provide visual context without spoiler, we utilize the previous chunk (4 seconds prior) as input.

Explanation Explanation entails inferring the most plausible psychological interpretation of an observed cue within context.

Next-Utterance Prediction Predicting the next line of dialogue following a marked cue serves as a proxy for inferring mental state capabilities. We input a marked transcript excerpt and speaker of the utterance, and output the utterance. Distractors are randomly sourced from the script corpus with semantic distances.

5.2 Input modality

Visual Cues: Frames The models receive visual token inputs consisting of a series of images representing moments when NVCs occur during a dialogue. Given that 32 frames would consume a significant number of visual tokens, we dynamically adjust the image size to a minimum of 64 to avoid exceeding the context length limits of vision-language models.

Text Cues: Script and Vocal Cues We also provide the script along with annotated vocal cues as input. The input consists of 1-minute script segments paired with truncated video clips.

5.3 Results

Clear Human-AI Gap In Table 3, even non-experts human demonstrate superior performance compared to the best AI model in Detection and Explanation task. Within same group, clear scale effect that large version model outperforms smaller model with big gap.

Larger Models Excels in Explanation over Cue Generation Excluding Gemini-Flash-1.5, larger models such as Qwen 2.5-32B and Qwen 2.5-7B exhibit notably stronger performance in Explanation tasks compared to Cue Detection (32B: 65.5% vs 47.7%, 7B: 65.1% vs 46.8%). They also show superior performance in context-dependant tasks such as Prediction. This trend suggests that the

context comprehension capabilities of these larger models may facilitate more accurate psychological interpretations.

Detection Binary vs Detection MCQ In the binary detection task, models generally perform better than in the multi-choice (MCQ) detection setting, suggesting lower cognitive load and ambiguity in label selection. However, some smaller models such as InternVL3-2B shows extreme gains in Binary task (95.6%), highlighting that model with high recall is advantageous, as all the label is 'Appears'.

Struggles in Invalid Understanding In *Explanation* task, the ability to discern invalid cues remains a significant challenge across all models. Even the strongest model, GPT-4o, struggled with high rates of false alarms in the invalid category (Invalid 50.6% vs Valid 64.9%).

5.4 Over-interpretation vs. Under-interpretation

In Figure 4 and Table 4, we categorize the model answer and the groundtruth combinations in *Explanation* task.

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: 'Valid' is a specific psychological meaning (*e.g.*Stressed). We define False Negative (FN) and False Positive (FP) as under-interpretation and over-interpretation.

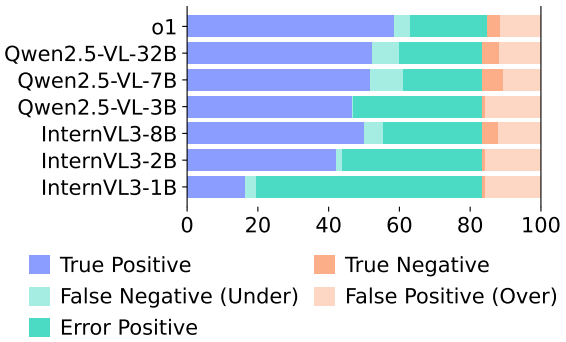

Figure 4: Stacked bar plots of LLM *Explanation* task answers.

The larger vision-language models—Qwen 2.5-VL-32B and InternVL3-8B—achieve the highest counts of both true positives and true negatives, and thus the best overall accuracy.

When we break down the errors (False Negatives + False Positives + Error Positives), a clear size effect emerges: as we move to smaller models, the share of EP (Error Positives) increases sharply.

Smaller models make more over-interpretation mistakes on the 'Invalid' labels, and when we concentrate on the 1B–3B models, they almost always assess the given NVCs as valid. Therefore, it appears that they have almost no ability to respond with "invalidity" according to the context in which the nonverbal cue occurs.

Finally, when we compare over-interpretation and under-interpretation, we can see that over-interpretation occurs far more frequently. Despite a label imbalance where valid ground-truth is far more numerous than invalid ground-truth, the models almost never commit under-interpretation, whereas over-interpretation accounts for most of the invalid-cue inputs.

5.5 Qualitative results

Figure 5 representative cases where the O1 model produces incorrect inferences in Detection-binary (first row) and Explanation (second row) tasks. In Detection-Binary task, the model misidentifies even clear cues such as 'neck touching' 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 with validated dictionary to cover comprehensive range of body language.

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 focuses on grounded, multimodal social-reasoning chains, but our MOTION2MIND isolates pure 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 the comprehensive evaluation framework with benchmark 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

Coverage of Nonverbal Behaviors While our dataset incorporates a comprehensive range of NVCs from established literature, it cannot exhaustively capture the full spectrum of human nonverbal communication. Cultural variations in gesture interpretation, micro-expressions, and complex combinations of simultaneous nonverbal signals remain challenging to represent fully in our framework. Additionally, our reliance on a single body language dictionary, though expertly curated, may not capture emerging or culturally specific nonverbal behaviors.

Simplified Assumptions in Action Recognition Our framework assumes perfect detection of NVCs in both text and video modalities, which may not reflect real-world challenges in action recognition. While this assumption allows us to focus on evaluating higher-level understanding, it potentially oversimplifies the complexities of detecting subtle movements, continuous motion, and overlapping gestures in practical applications. Future work should address the integration of actual action recognition systems and their associated errors.

Limitations of Synthetic Data Although our synthetic data generation approach enables a systematic evaluation of edge cases, it may not fully capture the naturalness and spontaneity of human nonverbal communication. The use of GPT-4o for data generation, although carefully controlled, could introduce biases or artifacts that differ from natural patterns of nonverbal behavior in human interactions.

9 Ethical Considerations

Privacy and Consent While our video dataset uses publicly available movie 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 potentially 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

Speaking rate—tools for clear speech. https://tfcs.baruch.cuny.edu/speaking-rate/. Accessed 2025-05-19.

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*.

Luigi Anolli, Rita Ciceri, and Maria Grazia Infantino. 2012. Markers of deception in italian speech: Pitch, response latency, and speech rate. *Phonetica*, 69(4):197–211.

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.

Ralph R. Behnke and Chris R. Sawyer. 2001. Public speaking anxiety as a function of trait anxiety and physiological reactivity. *Communication Research Reports*, 18(2):137–145.

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*.

Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

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*. Routledge, New York.

Dana R. Carney, Amy J. C. Cuddy, and Andy Y. 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*.

Clipzone Sitcoms. 2025. Clipzone sitcoms youtube channel. https://www.youtube.com/@ClipzoneSitcoms. Accessed: April 24, 2025.

Jesse I. Davis and Ann Senghas. 2010. Natural emotion elicitation and facial actions. *Emotion*, 10(6):862–874.

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. Face and emotion. *American Psychologist*, 48(4):384–392.

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

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.

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.

Friends. 2025. Friends official youtube channel. https://www.youtube.com/@Friends. Accessed: April 24, 2025.

David B. Givens. 2016. The nonverbal dictionary of gestures, signs, and body language cues. https://center-for-nonverbal-studies.org/. Accessed 2025-05-19.

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

Ursula Hess and Patrick Bourgeois. 2010. You smile—I smile: Emotion expression in social interaction. *Biological Psychology*, 84(3):514–520.

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.

joblo. 2025. Joblo movie clips.

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

Keeping Up with the Kardashians. 2025. Keeping up with the kardashians official youtube channel. https://www.youtube.com/@KUWTK. Accessed: April 24, 2025.

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.

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

Michael W. Kraus. 2019. Self-touch as a regulator of social contact. *Social Psychological and Personality Science*, 10(4):479–486.

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*.

lionsgate. 2025. Lionsgate movies.

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*.

Abbie Marono, David D Clarke, Joe Navarro, and David A Keatley. 2017. A behaviour sequence analysis of nonverbal communication and deceit in different personality clusters. *Psychiatry, Psychology and Law*, 24(5):730–744.

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. Reading facial expressions of contempt. *Personality and Social Psychology Bulletin*, 34(5):734–745.

Albert Mehrabian. 1972. *Nonverbal Communication*. Aldine-Atherton, Chicago.

Joe Navarro. 2018. *The dictionary of body language: a field guide to human behavior*. HarperCollins.

Joe Navarro and John R Schafer. 2001. Detecting deception. *FBI L. Enforcement Bull.*, 70:9.

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.

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.

Klaus R. Scherer, Harvey London, and Jared J. Wolf. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7(1):31–44.

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*.

The Office. 2025. The office official youtube channel. `https://www.youtube.com/@TheOffice`. Accessed: April 24, 2025.

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.

Rose Tramposch, William DeJong, and Jessica L. Tracy. 2021. Postural expansion and emotional expression jointly signal pride. *Emotion*, 21(4):969–980.

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.

Aldert Vrij and Rob Taylor. 2004. Response latency as a cue to deception. *Legal and Criminological Psychology*, 9(2):159–181.

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.

Justin W. Weeks, Chao-Yang Lee, Alison R. Reilly, Ashley N. Howell, Christopher France, Jennifer M. Kowalsky, and Ashley Bush. 2012. "the sound of fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. *Journal of Anxiety Disorders*, 26(8):811–822.

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

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.

Mao Yamamoto, Akihiro Takamiya, Kyosuke Sawada, Michitaka Yoshimura, Momoko Kitazawa, Kuo-Ching Liang, Takanori Fujita, Masaru Mimura, and Taishiro Kishimoto. 2020. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE*, 15(9):e0238726.

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*.

A More Analysis

A.1 Knowledge: Validity-binary task

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 5: Valdity binary task results. Random score is 0.5.

Besides *cue* and *explanation* task in §3, We measure performance of *validity-binary* task, which is to guess if the combination of nonverbal cue and psychological meaning is valid or not (*e.g.*Input: Arm crossing - self-protection / Output: True).

We source the positive sample in the dictionary, and match the negative sample with semantic distance between *True* explanation pool of the cue. Basic random score is 0.5.

In Table 5, we see the model accuracy, precision and recalls are all generally high except 0.5B models. Even 1.5B models show above 0.8 score for three metrics. This implies high level of knowledge about Nonverbal cue and possible meanings.

A.2 Categorical Performance Difference

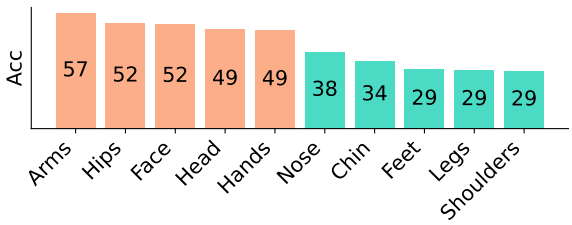

Figure 6: 5 most accurate (Orange) and inaccurate (Green) body parts. Models are less likely to choose 'invalid' responses when similar NVC is added to the dialogue (x: NVC numbers, y: Answer as invalid) for both validity and explanation tasks.

Contrary to common assumptions that facial cues—often considered the 'window to the mind'—would yield the highest accuracy in nonverbal cue interpretation, while arms, hips, and hands/fingers exhibit relatively higher accuracy. This unexpected pattern suggests the process of translating these cues into verbal descriptors (e.g., 'eye darting') can introduce ambiguity and semantic drift, complicating accurate recognition. Furthermore, lower accuracy in regions such as legs, nose, and feet may reflect their less frequent role as focal points in everyday nonverbal communication.

A.3 Appearing Human size

To determine whether recognition accuracy varies with the on-screen size of a person, we computed the correlation between the average area of the human bounding box and the model's accuracy. The analysis revealed virtually no association—the overall mean correlation coefficient was –0.005 ± 0.065—indicating that bounding-box size has little influence on accuracy.

A.4 Frame Numbers

We set the max frame numbers as 16, and we measure the performace change based on the input frame numbers. We uniformly sample frames at equal intervals from each GIF (up to a maximum of 32 frames). In Figure 7, as the number of frames decreases, the non-verbal cue motions become harder to discern, and the model's performance correspondingly declines. *Explanation* shows minor drop than *Detection* as it use script as another information to answer question.

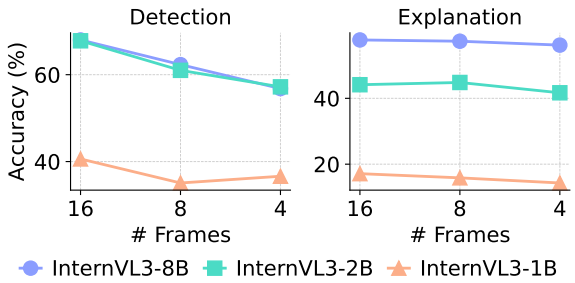

Figure 7: Performance change based on max frames number. In §5, we set max frames to 16.

B Inference Details

For inference, we utilize maximum 4 NVIDIA GeForce RTX 3090 for inference especially for 32B Video-language Model. For other 8B, 7B, 3B, 2B, 1B we adopt 1 3090 GPUs. Using paged attention served in VLLM library (Kwon et al., 2023), inference for one task type takes less than 2 hours.

For hyperparamenters, we utilize 0 temperature for reproducibility, seed 0, max tokens 500, top p 0.001, repetition penalty as 1.05.

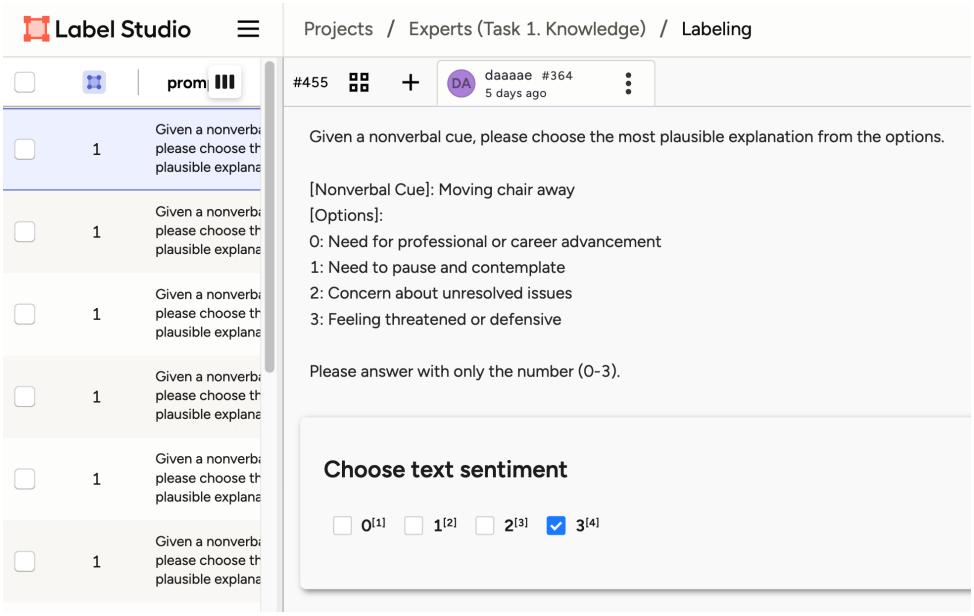

Figure 8: Example of the labeling interface.

C Human Evaluation

C.1 Annotator Selection

Experts We recruited 3 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 Studio[2] using the interface shown in Figure 8.

D 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)

E MOTION2MIND

E.1 Comprehensiveness of mind state labeling

In Table 6, we list up the representative mind state labels which can be categoried by the 6 mental states described in Ma et al. (2023).

In Figure 9, we visualize treemap to see the most frequent 15 motion cue labels with 5 most frequent mind state labels for each.

E.2 Cultural Universality

Our source dataset is derived from Navarro's body-language dictionary (Navarro, 2018), which catalogues nonverbal cues that are widely considered biologically based and cross-culturally universal

[2] https://labelstud.io/

reflections of internal mental states. Examples include eyebrow flashes or raised brows (affiliation/-greeting), eye-widening (surprise or fear), prototypical anger and sadness expressions, jaw drops, the disgust-related nose-wrinkle (bilateral or unilateral), nostril flaring, teeth baring, blink-rate acceleration, and pupil dilation. Because these signals are largely automatic and tied to autonomic physiology, their form and interpretation remain remarkably stable across societies.

By contrast, gestures such as head nods and shakes, the thumbs-up or 'OK' sign, and norms governing direct eye contact are strongly culture-bound—their meanings can even reverse from one region to another. Mapping this culture-specific layer of nonverbal communication is a valuable research problem in its own right, but it lies beyond the scope of the present study.

F Psychological Grounding

F.1 The dictionary of body language

Joe Navarro is a behavioral analysis expert who served in the FBI for over 25 years, and his book Navarro (2018) is widely cited in psychology and rhetoric. Based on his extensive experience, he compiles 407 reliable NVCs (nonverbal cues). His research (Navarro and Schafer, 2001) has been extensively used in FBI and police investigations, while studies have shown the correlation between unconscious body signals and psychological states (Marono et al., 2017).

His 407 cue entries span every major body region—*Eyes, Neck, Nose, Face, Hands/Fingers, Cheeks/Jaw, Chest Torso, Belly, Hips Buttocks (Genitals), Chin, Eyebrows, Arms, Forehead, Shoulders, Mouth, Feet, Head, Legs, Lips, Ears*. Below we highlight representative scientific findings that mirror Navarro's field claims and justify our choice to adopt his taxonomy.

- **Eyes.** Pupil dilation tracks interest and arousal (Bradley et al., 2008). Direct gaze signals confidence but, in hostile settings, dominance (Kleinke, 1986).
- **Nose.** Nostril-flare marks anger/high arousal, while brief nose-wrinkle pairs with disgust (Ekman, 1997).
- **Mouth & Lips.** Lip compression indexes tension or withheld opinion (Ekman, 2003). A lateral lip-purse conveys disagreement (Hess and Bourgeois, 2010). One-sided lip raise (AU 14) universally signals contempt (Matsumoto and Hwang, 2008).
- **Cheeks & Jaw.** Clenched jaw correlates with anger or restraining speech (Burgoon et al., 2016). Cheek sucking (drawing cheeks inward) often precedes decision anxiety (Navarro, 2018).
- **Forehead & Eyebrows.** Brow-lower (AU 4) denotes anger/effort; raised inner brows (AU 1+2) indicate fear or pleading (Ekman, 2003). Forehead tension peaks during concentration or stress (Davis and Senghas, 2010).
- **Arms & Hands.** Crossed arms predict defensive or closed attitudes (**?**). Arms-akimbo (hands on hips, elbows out) broadcasts dominance and confidence (Carney et al., 2010). Self-touch (arm rubbing, neck stroking) functions as a pacifier under anxiety (Kraus, 2019).
- **Shoulders & Torso.** Shoulder shrug (elevated, rotated) conveys uncertainty; shoulder slump co-occurs with sadness and low confidence (Mehrabian, 1972).
- **Chest / Belly.** Expansive chest displays pride or high status, whereas inward chest and belly guarding indicate vulnerability (Tramposch et al., 2021).
- **Hips, Buttocks, Genitals.** Pelvic retreat (hips angled away) shows discomfort; pelvis forward with relaxed stance signals attraction or confidence (Givens, 2016).
- **Legs & Feet.** Sudden leg uncrossing or weight-shift marks threat arousal. Increased foot fidgeting accompanies nervous energy.

Collectively, these peer-reviewed findings validate Navarro's cluster-based approach and demonstrate that his dictionary encompasses empirically supported NVCs across the full spectrum of body parts.

F.2 Vocal Cue Annotation

Our algorithm maintains a rolling baseline for each speaker (last 50 segments $\approx$ 5 min) and tags utterances when *(i)* words-per-minute (WPM) or *(ii)* fundamental frequency (F_0) exceed the speaker's mean by $+1\sigma$ (`[FAST]`, `[HIGH_PITCH]`), or *(iii)* silence lasts $\geq$ 600 ms and covers $\geq$ 5% of the

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 6: 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.

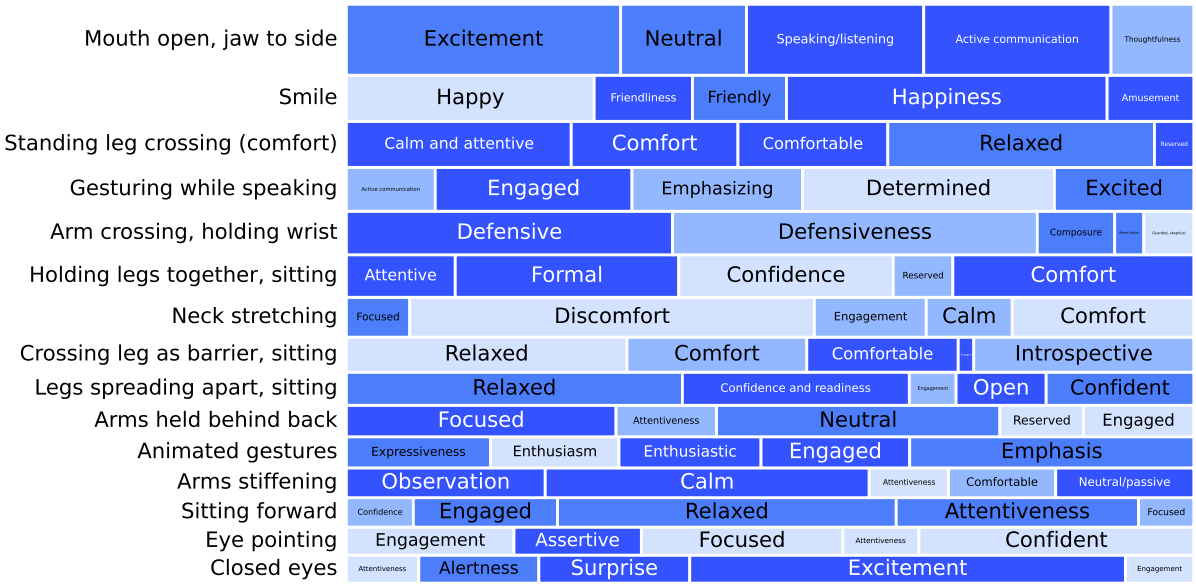

Figure 9: Treemap of top 15 most frequent nonverbal cues and its psychological explanation. We clustered the explanation with the semantic embedding and title as one of each cluster.

segment ([LONG_PAUSE]). Segments shorter than 120 ms skip F_0 analysis, and baselines update after each decision.

Pitch. Social-threat and anxiety tasks reliably raise F_0 by roughly one standard deviation above baseline (Weeks et al., 2012), and meta-analytic work shows a comparable pitch lift during deception (Anolli et al., 2012). Conversely, major depression produces lower, flatter pitch contours (Yamamoto et al., 2020). Labeling $F_0 > \mu + \sigma$ therefore captures the arousal-linked excursions of interest while respecting individual vocal ranges.

Speech rate. Average conversational English hovers around 150 WPM (tfc); state-anxious speakers often surge beyond 180–200 WPM—about $+1\sigma$ on personal baselines (Behnke and Sawyer, 2001). In contrast, depressive speech frequently drops below 110 WPM (Yamamoto et al., 2020). A $+1\sigma$ rule thus isolates the adrenaline-driven accelerations without penalising naturally brisk talkers.

Pauses. Depressed individuals and deceivers show markedly longer within-utterance pauses and response latencies (often > 600 ms) compared with controls (Yamamoto et al., 2020; Vrij and Taylor, 2004). Confident speakers, by contrast, use brief, infrequent pauses (Scherer et al., 1973). Tagging silences that both exceed 600 ms *and* occupy $\geq 5\%$ of the segment pinpoints these cognitively-loaded gaps.

Rolling z-scoring is recommended in vocal-affect research to remove inter-speaker physiological variance while highlighting state deviations (Weeks et al., 2012). Hence, our adaptive thresholds translate decades of paralinguistic evidence into objective, speaker-normalised labels.

G 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.

H Prompts

In Table 7 and Table 8, we specify the prompts we use for §4 and §5.

I 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 $\leftarrow$ **true**$^{|I|}$

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

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 $\mathcal{A} \leftarrow \{j \mid C[j] \neq t.\mathit{pivot}\}$
Greedy selection
$\mathcal{S} \leftarrow [\,]$

while $|\mathcal{S}| < k$ **do** dir = far pick $j^\star = \arg\min_{j \in \mathcal{A}} \sigma[j]$ pick $j^\star = \arg\max_{j \in \mathcal{A}} \sigma[j]$
$\mathcal{S} \mathrel{+}= [\,I[j^\star]\,]$; $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j^\star\}$
Assemble MCQ entry
$\mathcal{R} \mathrel{+}= \langle t, [\,t\,] \cup \mathcal{S} \rangle$ **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 7: 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 8: 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.