Wait... Was That a Sign? 🤖 Reading Minds Through Actions: OBSERVABLE THEORY OF MIND with Nonverbal Cues

Anonymous ACL submission

Abstract

Our ability to interpret others' mental states with nonverbal cues (NVC) has been 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 OBSERVABLETOM, a comprehensive framework for evaluating the ToM capabilities of machines in interpreting NVCs. Starting from an FBI agent's validated profile handbook, we develop OTOM$_{text}$, a dialogue dataset of 9,896 entries with diverse context, and OTOM$_{video}$, a carefully curated video dataset with fine-grained annotations of actions with psychological interpretations. Our evaluation reveals that current AI systems struggle significantly with NVC interpretation, showing not only a substantial performance gap (GPT-4o: 73.6% vs. human: 91.5%) but also patterns of over-interpretation, with particularly low precision (40.0-63.5%) indicating high false alarm rates.

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 even 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 the subtle mental states expressed through nonverbal communication remains largely unverified. Although existing Theory of Mind (ToM) benchmarks (Le et al., 2019; Weber et al., 2021; Jin et al., 2024) have advanced our understanding, they primarily focus on false-belief tasks (Wimmer and Perner, 1983) - testing an agent's ability to reason about asymmetric information between characters. However, human social cognition encompasses a much broader spectrum of mental state inference (Ma et al., 2023), involving the dynamic exchange of NVCs.

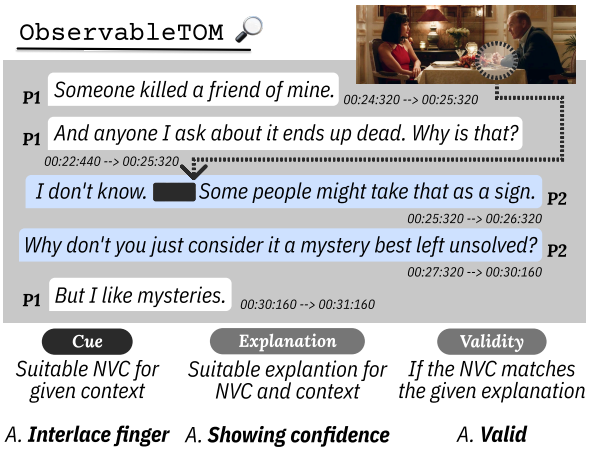


Figure 1: We build OBSERVABLETOM, a dataset designed to integrate nonverbal cues (NVC) understanding in context.

Another attempt to measure NVC understanding capability through video datasets (Luo et al., 2020; Chen et al., 2023; Liu et al., 2021; Huang et al., 2021) have encountered two significant methodological limitations. First, they employ oversimplified scoring systems 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 with its psychological interpretation - for instance, identifying which exact moment in a

video sequence indicates that a subject is 'happiness' or 'proud of themselves'.

To address these challenges, we introduce OBSERVABLETOM (OTOM), a comprehensive framework for evaluating machines' ToM capabilities in interpreting nonverbal social cues. Our framework starts from an expert-established psychological literature about NVCs. We expand that theory with detailed everyday NVCs in detailed dialogue generated with GPT-4o (OTOM$_{text}$) or grounded in movie clips (OTOM$_{video}$). Our data is validated by a high score of human labelers showing its plausibleness and clarity. While current state-of-the-art model GPT-4o (OpenAI et al., 2024a) correctly guess complex false belief tasks, they fail to understand day-to-day NVC in real-world simulating contexts.

Our key contributions are:

1. OTOM$_{text}$, specifically designed for evaluating nonverbal ToM capabilities, featuring fine-grained behavioral cues and their corresponding mental states.

2. OTOM$_{video}$, A NVC benchmark sourced from movie clips to simulate multi-model real-world NVC understanding.

3. Empirical analysis demonstrating significant gaps between the human-like social agent and current LLM/VLM.

In §2, we examine key challenges in theorizing NVCs. §3 evaluates basic nonverbal understanding capabilities without contexts. §4 introduces our OBSERVABLETOM framework, and §5 and §6 presents empirical analyses of current models.

2 What Makes Understanding NVCs Difficult?

2.1 Ambiguous Definition of NVCs

Challenge Defining NVCs presents two fundamental challenges. First, determining *what are meaningful nonverbal signals* is an ongoing debate in psychology. As human rarely stand still, conceptualizing what behavior is NVC is important. Second, mapping these physical signals to the underlying psychological states is also ambiguous. For example, research on lip biting behavior shows conflicting interpretations: some studies link it to anxiety and emotional suppression (Zuckerman et al., 1981), while others suggest it could be a habitual action without psychological significance (Harrigan, 2005).

In OTOM To address this definitional challenge, our work leverages a body language dictionary (Navarro, 2018), which was written by a former FBI Agent with decades of field experience. As illustrated in Figure 2, NVCs deal with the whole body in a comprehensive way. It also offers multiple interpretations of NVCs (e.g., Be stressed, be happy, or to show status) for one cue. We disentangle the explanation paragraph with GPT-4o. The outer circle displays the five most frequent interpretations for each category.

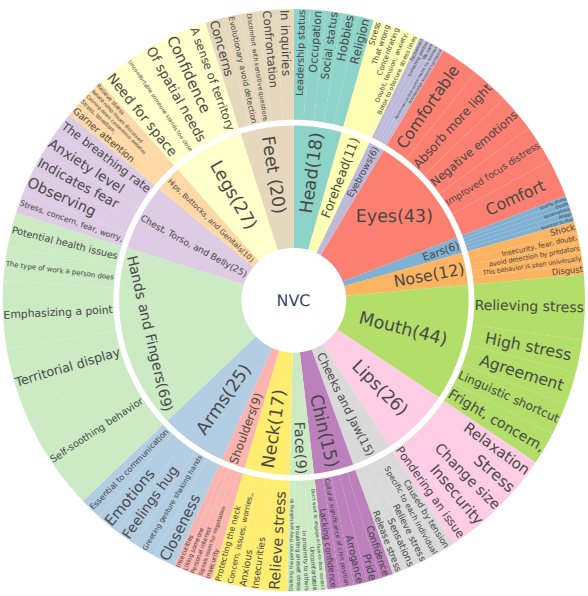


Figure 2: Anatomical categories in data source (Navarro, 2018) and the five most frequent interpretations for each category. It includes 407 NVCs, and we get 1,114 interpretations by disentanglement (the prompt is provided in Table 10). The wider plot is in Figure 8.

2.2 Contextual Defeasibility

Challenge Since 'real mind' of other agent is inapproachable, all interpretations exist as possibilities. In that case, context emerges as a crucial influencing factor in the interpretation of NVCs. Humans naturally excel at adjusting their interpretations of NVCs based on contextual shifts. For LLMs to achieve meaningful human interaction, they must develop the capability to accurately incorporate contextual understanding into their interpretation of nonverbal expressions.

In OTOM To systematically address this context dependence, we employ Dell Hymes' SPEAKING model (Hymes, 2013) as our analytical framework.

Model	Cue	Explanation	Validity		
	Acc	Acc	Acc	Pre	Rec
Random	25.0	25.0	50.0	50.0	50.0
GPT-4o	**74.3**	**86.2**	**82.4**	**70.5**	92.8
GPT-4o-mini	72.0	84.0	79.4	63.2	93.5
Qwen2.5-32B	71.2	83.3	79.6	63.9	93.1
Qwen2.5-14B	67.0	82.1	74.3	51.4	**94.7**
Qwen2.5-7B	66.5	75.6	65.2	32.4	94.0
Qwen2.5-3B	62.4	75.0	65.1	32.4	93.8
Qwen2.5-1.5B	62.4	75.0	65.1	32.4	93.8

Motion	*Neck Streching*
Dictionary	...in a circular motion is a **stress reliever** and **pacifier**. This is often seen when people are asked difficult questions they would rather not answer.
GPT-4o	1. **Relief or Release of Tension**: to relieve tension or stress... 5. **Confidence or Dominance**: ... as it exposes a vulnerable part of the body...
Qwen2.5-32B	In a professional or formal setting, it might indicate **discomfort**, **tension**... **physical stiffness** or ...
Qwen2.5-14B	It can indicate **physical discomfort** or **tension**, ...it might also be a preparatory gesture,...

Table 1: (Left) Nonverbal cue (NVC) knowledge scores without context, tested using a validated source (Navarro, 2018). We measure accuracy (Acc), precision (Pre), and recall (Rec) to assess validity. (Right) Real examples illustrating the psychological interpretation of *neck stretching*.

SPEAKING has been validated and utilized by numerous subsequent studies (see Appendix C). The framework encompasses crucial components that can sophisticate communication meanings: Setting and Scene (S), Participants (P), Ends (E), Act Sequence (A), Key (K), Instrumentalities (I), Norms (N), and Genre (G).

2.3 Limitations of Action Recognition

Challenge Even for the same NVC (e.g., head nodding), there exist significant interpersonal and intrapersonal variation in intensity. For frequency-dependent NVCs (e.g., eye blinking), differences in frequency can occur.

In OTOM As there can be various modalities and detection methods such as video language understanding, motion captioning, and sensor data understanding in real-world applications, we suppose NVC detection is given. In our OTOM$_{text}$ and OTOM$_{video}$, we incorporate NVCs between verbal cues, similar to stage directions in the screenplay format. In OTOM$_{video}$, we also pinpoint the NVC-appearing frames and ask VLMs to explain the cue and infer its psychological interpretations.

3 Basic Capability: Cue Understanding Without Context

We evaluate LLMs' performance with our validated source (Navarro, 2018) to measure their exactness of knowledge with minimal effect of context.

3.1 Methods

Data As mentioned in §2, we structure the dictionary into a 1 cue:n explanations, which allows us to test three types of understanding: (1) *Cue*: identifying the most appropriate NVC for a given explanation, (2) *Explanation*: selecting the most fitting explanation for a given cue, and (3) *Validity*: validating whether a specific cue-explanation pair represents a legitimate connection.

Multi-choice QA We design multiple-choice questions where several explanations could be valid for a single NVC. Using semantic embeddings[1], we deliberately select distractor options with semantically distant *explanations* from the *explanation* of correct answer. For example, while 'crossed arms' might indicate both 'defensiveness' and 'comfort-seeking', the model must distinguish these valid interpretations from semantically unrelated but plausible-sounding options like 'expressing excitement'. For *Cue* and *Explanation* tasks, models select from four distinct options, while for *Validity*, they choose between *valid* and *invalid*.

3.2 Results

Good score, with larger the better. GPT-4o outperforms all Qwen2.5 models in validity, precision, and recall metrics, with consistent scaling benefits observed in the Qwen2.5 series from 7B to 32B. This demonstrates that larger model shows better theoretical understanding of NVCs, particularly evident in precision scores (32.41 $\rightarrow$ 63.91).

Choosing best cue is difficult than best explanation. Models show higher proficiency in generating appropriate explanations for given cues compared to identifying relevant cues for situations. This implies that suggesting the contextual implications of nonverbal behaviors is more challenging than providing explanations for pre-identified cues.

Low precision: Frequent false positive. All models demonstrate significantly higher recall than

[1]All semantic embeddings in this paper use all-MiniLM-L6-v2 (Reimers, 2019).

precision scores in binary *Validity* task, with extreme cases like Qwen2.5-7B (Yang et al., 2024) showing a stark contrast (32.41 precision vs 94.01 recall). This indicates that models tend toward over-interpretation, often generating false positives, suggesting a need for more conservative classification thresholds.

4 OBSERVABLETOM: Multimodal NVC Dataset with Defeasible Contexts

We build OBSERVABLETOM, a testbed about NVC understanding of LLM, Vision language models (VLM), and VideoLM in defeasible contexts. It consists of OTOM$_{text}$ (§4.1), a synthetic text dataset of 10k entries, and OTOM$_{video}$, a meticulously annotated multimodal video dataset of 200 entries (§4.2).

4.1 OTOM$_{text}$

Why Synthetic Dataset? Creating a large-scale annotated NVC dataset poses several challenges. First, capturing the full spectrum of NVCs in natural settings is resource-intensive, particularly in rare but significant cases (Knapp et al., 1978). Second, NVC interpretation without any grounded concept can be highly subjective by annotators (Harrigan, 2005). To address these challenges, we leverage GPT-4o and validate with human accuracy to generate a comprehensive dataset.

Source 1: The dictionary of Body Language (*Cue*, *Explanation*) First, we establish a foundational understanding of NVCs using *The Dictionary of Body Language*, which provides 1,114 pairs of NVCs curated by the expert. These pairs serve as ground truth for a consistent NVC interpretation.

Source 2: 'Invalidity' integration with SPEAKING Framework Since humans frequently exhibit habitual and unconscious NVCs (Yager et al., 2009), we include 'invalid' cues in our dataset. We leverage the SPEAKING framework to ground the validity of NVCs, generating 10 diverse scenarios for one pair of NVCs and possible explanation. The first scenario establishes a highly probable context (e.g., hand trembling, nervousness, during a first date), while the remaining nine systematically vary contextual factors to create scenarios where the same NVC becomes psychologically irrelevant. For example, the same hand-wringing gesture might be a habitual motion while watching television. Through human validation, this context-based approach achieves an accuracy of 91.5% in distinguishing valid from invalid cue interpretations (see Appendix A). We exclude the 'Norm' element from the SPEAKING framework, as cultural variations would require factual grounding beyond our current scope.

Content
Cue Ventilating neck
Explanation Relieve discomfort caused by stress
Context An office worker in a high-pressure meeting room, surrounded by colleagues and superiors.
More cues Eyes closed, rubbing bridge of nose, Neck massaging, Pulling clothing to ventilate
Dialogue Ryan: (adjusts collar) Alright, team, we need to rethink our approach for the next phase. The client's not happy with our current trajectory. [__] Laura: I agree. We need fresh ideas. Maybe we focus more on social media engagement this time? [Neck massaging] Sam: I can put in extra hours to help with the analysis, though I might need some guidance on the new strategy expectations. [Pulling clothing to ventilate] Megan: (closes her eyes briefly) I just received an email from the client with some feedback. It's... not very positive. [Eyes closed, rubbing bridge of nose] Ryan: Thanks, Megan. We'll need to address their concerns point by point. Sam, your extra effort is noted and appreciated. Laura: (leans forward) Can we schedule a brainstorming session tomorrow? I'd like to have more diverse input before we finalize anything. Sam: That sounds good, Laura. I'll prepare some data points that might assist during that session. Megan: (taking a deep breath) I'll compile the client feedback and distribute it to everyone, so we're all on the same page.
Options 0: Scarred ears 1: Elbows spreading out 2: Cheek framing **3: Ventilating neck**

Table 2: An example of OTOM$_{text}$. It is a dialogue including multiple nonverbal cues (with their possible psychological interpretation) and defeasible contexts that can either validate or nullify the NVC's interpretation.

Dialogue with Multiple Cues (*More cues*) We incorporate both valid and invalid cues for each dialogue to construct dialogues containing multiple NVCs. We cluster 3-6 cues from the context pool, with semantic similarity of context, to maintain natural dialogue flow. Our empirical analysis determines that 3-6 produces optimal prompt following when generating dialogue; larger clusters result in unnatural dialogue flow or deviate from

our specified parameters.

Dialogue Generation (*Dialogue*) We utilize GPT-4o to generate naturalistic dialogues incorporating the clustered NVCs and their contexts. The generation process is constrained by three rules: (1) The source NVC should appear once (we control the appearance as a factor in §5.4), (2) adherence to the SPEAKING framework conditions (*valid* or *invalid*), and (3) subtle and natural integration of cues without explicit reference to their interpretative meanings.

Filtering and Postprocessing Our automated postprocessing pipeline implements two primary filtering mechanisms: elimination of redundant NVC instances (retaining only the first occurrence) and removal of generations that failed to incorporate the specified NVC. To ensure scalability, we opt for prompt optimization through iterative manual inspection rather than human filtering. When evaluating against OTOM$_{text}$, this approach achieves reliability scores of 83.3% for cue validity and 73.3% for explanation.

MCQ Generation (*Options*) Similarly with §3, we construct multiple-choice questions by selecting maximally divergent explanations based on semantic embedding. This evaluation framework accommodates multiple valid interpretations of a single NVC, assessing the model's ability to identify the most contextually appropriate explanation among semantically distant, nonsensical alternatives. With this pipeline, we get 9.8k dialogue dataset with annotation of multiple NVCs and its psychological meaning.

4.2 OTOM$_{video}$

Sourcing Movie Clips We source 797 movie clips from YouTube channels (lionsgate, 2025; joblo, 2025) specialized in movies. We only use 'Official Clip' as it has less scene transition. To generate subtitles, we perform speech-to-text conversion with Whisper-large-v2 (Radford et al., 2022) and speaker diarization with Pyannote (Bredin et al., 2019) and combine the information according to the timestamp.

Filtering and Annotation We make OTOM$_{video}$ from 200 movie clips, each containing 6-35 lines of dialogue. This range is chosen based on our empirical observation that shorter clips lack sufficient verbal context for meaningful analysis.

Figure 3: An example of a nonverbal cue in OTOM$_{video}$. We annotate NVCs with their corresponding mind state interpretations that appear within 32 image frames in the videos.

The preparation of the dataset involved: (1) manual correction of STT and speaker diarization errors and the addition of crucial multimodal context in parentheses (e.g., Scene Changed, Running), (2) balanced genre distribution to capture diverse and natural emotional expressions, and (3) manual annotation of NVCs using the predefined dictionary, resulting in 167 unique NVCs and 185 psychological explanations. Each clip contains 1-4 annotations, with NVCs deliberately distributed throughout the dictionary to ensure comprehensive coverage.

Frame Extraction Based on research (Birdwhistell, 1970) showing that human NVC typically occurs in brief interaction sequences, for the visual component, we extract frames at 8 FPS within 4-second windows around each annotated NVC (a maximum of 32 frames). This window size is determined to be optimal for capturing the complete temporal context of human NVC while maintaining dataset efficiency. The extracted sequences are all verified with manual checking steps.

	Items	Cues	Source	# Emotions	Actions
OTOM$_{text}$	9.8k	5.6	GPT-4o	1k	✓
OTOM$_{video}$	200	1.6	Movie	185	✓
Aff-Wild2	548	-	Interview	10	✓
iMiGUE	359	5	Interview	2	✓
BoLD	10k	1.3	Movie	28	✗
THEATER	258	1	Movie	8	✗

Table 3: Overview of OBSERVABLETOM, which consists of OTOM$_{text}$ and OTOM$_{video}$, providing rich cues and detailed explanations of emotional mind states and actions. Other benchmarks (Shafique et al., 2023; Liu et al., 2021; Luo et al., 2020; Kipp and Martin, 2009) lack either comprehensive mental state annotations or fine-grained action annotations.

To the best of our knowledge, OTOM$_{video}$ is the

	Model	Open	ToM Method	Cue	Explanation	Validity			
				Accuracy	Accuracy	Accuracy	Precision	Recall	F1
	Human	-	-	-	73.3	83.3	83.5	83.3	83.3
OTOM$_{text}$	GPT-o1	✗	Plain	37.5	30.2	56.8	40.0	95.7	56.4
	GPT-4o	✗	Plain	64.9	46.2	65.2	63.5	73.8	68.3
	GPT-4o-mini	✗	Plain	64.3	45.2	61.8	58.2	83.6	68.6
	Qwen2.5-32B	✓	Plain	65.0	44.5	66.0	61.0	85.0	71.0
	Qwen2.5-14B	✓	Plain	62.2	50.4	64.2	50.1	84.2	62.8
	Qwen2.5-7B	✓	Plain	63.9	44.9	56.7	41.7	79.9	54.8
	Qwen2.5-3B	✓	Plain	50.0	41.2	56.8	38.9	55.4	45.7
	GPT-4o	✗	Wilf et al. (2023)	91.2	50.5	64.4	51.8	81.8	63.4
	GPT-4o-mini	✗	Wilf et al. (2023)	91.2	50.5	64.4	51.8	81.8	63.4
	Qwen2.5-32B	✓	Wilf et al. (2023)	95.3	45.6	64.4	51.3	77.3	61.7

Table 4: Performance of LLMs on OTOM$_{text}$. LLMs generally perform worse than humans across Cue, Explanation, and Validity tasks. The random baseline is 25.0% since we provide four options for each question.

first multimodal video dataset that combines fine-grained action annotation with its psychological interpretation, grounded in naturalistic (cinematic) dialogue.

5 Test LLMs with OTOM$_{text}$

With OTOM$_{text}$, we test the ability of current LLMs to understand diverse contexts and modalities in a simulated environment.

5.1 Experimental settings

We follow the similar MCQ setting in §3 with some modifications: (1) For *cue*, we only utilize *valid* samples (2) for *explanation* task, *invalid* cue has 'No clue' option and it is the correct answer. (3) *Validity* now has 4 options: *highly valid*, *somewhat valid*, *somewhat invalid*, and *somewhat invalid* to accommodate the nuanced impact of contextual information.

We test with current high-performance models, and human with randomly sampled 500 instances for two graduate students (details in Appendix A).

5.2 Current LLMs' Performance

Much inaccurate than humans. In Table 4, the results demonstrate that while larger models such as GPT-4o (65.2%) show improved accuracy compared to smaller variants such as GPT-4o-mini (61.8%), there remains a substantial gap compared to human performance (83.3%). This intuitively back up the validity of OTOM$_{text}$, and it is notable that the powerful reasoning ability of GPT-o1 is not helpful in this task (56.8%).

Consistent poor performance The low explanation accuracy scores (ranging from 30.2% to 50.5%) and validity precision (40.0% to 63.5%) across all models suggest that they are fragile to false alarms in their predictions. This indicates that models often fail to appropriately identify when they lack sufficient information to make a valid inference - a crucial aspect of robust ToM reasoning.

Specialized Theory of Mind methods show limited improvements. The implementation of Wilf et al. (2023)'s ToM method shows modest gains in specific metrics, such as improving GPT-4o's cue accuracy from 64.9% to 91.2%. However, this method doesn't work for explanation or validity prediction, suggesting there are different mechanism in current information based ToM method and NVC understanding capability.

5.3 How do they deal with multiple cues?

As there would be a case dealing with multiple NVCs into consideration rather than one, we draw a scatter plot in Figure 4 to see correlation between *Signal-to-Noise Ratio* (number of valid NVCs in the dialogue) and *Semantic Heterogeneity* (average cosine distance of text embeddings).

Semantic Heterogeneity has a strong negative correlation with accuracy. This pattern is more evident in larger models (Qwen2.5-32B: -0.87, Qwen2.5-14B: -0.90). But larger model sizes consistently achieve better performance in handling heterogeneous tasks while maintaining higher accuracy levels. This is evidenced in both GPT-4 variants showing similar correlation coefficients (Y: -0.85), and in the Qwen2.5 models demonstrating

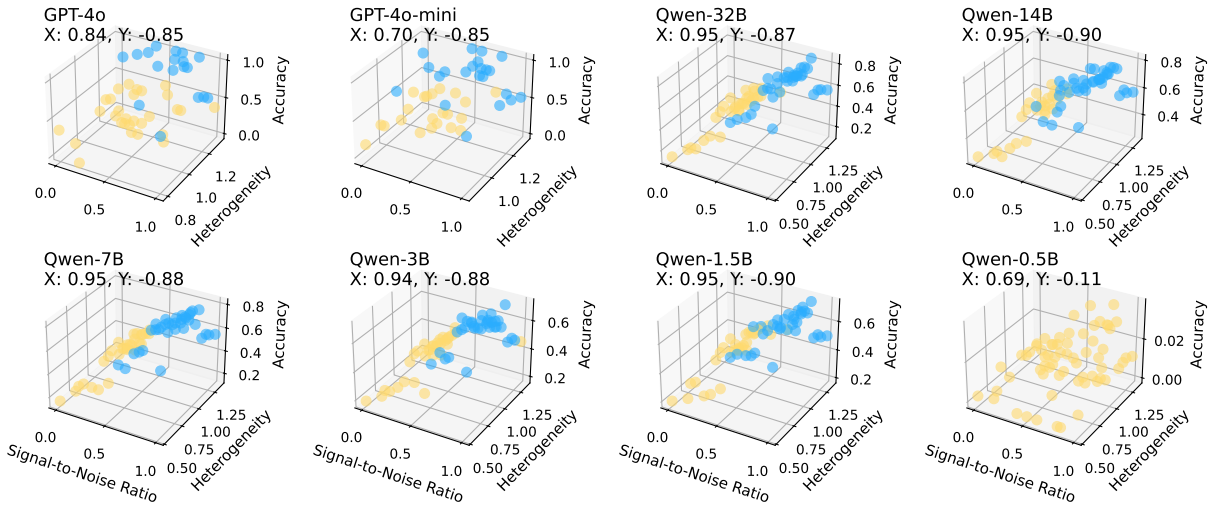


Figure 4: A 3D scatter plot showing Signal-to-Noise Ratio (X), Heterogeneity (Y), and Accuracy (Z), with points colored blue (accuracy > 0.5) and yellow (accuracy < 0.5).

systematic improvements from 0.5B to 32B (correlation strengthening from -0.11 to -0.87).

Signal-to-Noise Ratio Signal-to-Noise undermines other NVC understanding. The size of the model affects the signal-to-noise ratio performance differently. The larger models (Qwen2.5-32B, 14B, 7B) demonstrate better performance with higher signal-to-noise ratios (above 0.6) in high Signal-to-Noise conditions. In contrast, smaller models like Qwen2.5-0.5B show more scattered and less consistent performance patterns (around 0.02) and weak correlation (Y: -0.11).

5.4 Do LLMs prioritize verbal cue or nonverbal cue when conflict?

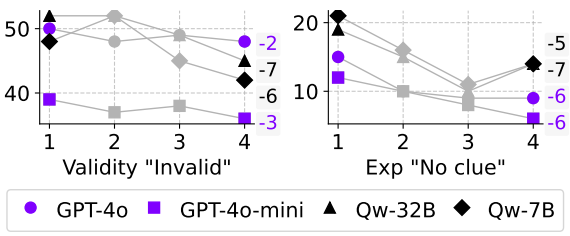


Figure 5: Models are less likely to choose 'invalid' responses when similar NVC is added to the dialogue (x: NVC numbers, y: Answer as invalid) for both validity and explanation tasks.

Just as humans consider comprehensive nonverbal cues (NVCs) to interpret individual NVCs, we examine 'invalid' cases where NVC and verbal cues show contradictions. To investigate which cues LLMs prioritize, we progressively introduce additional NVCs to dialogues that carry similar psychological implications to the existing NVCs.

Comprehensive Cue Processing In Figure 5, all models demonstrate a trend to consider NVC as 'valid' as additional peer NVC is added. This suggests that models integrate multiple communication cues when making validity judgments, rather than relying on individual cues in isolation. The pattern consistently appears across all model scales, from GPT-4o to Qwen-7B, indicating a fundamental capability in processing multiple cues.

Non-Linear Response Patterns The changes in model responses do not follow a linear trajectory, suggesting that models are not performing simple cumulative reasoning. Interestingly, the magnitude of these changes remains relatively consistent across different model scales, indicating that the ability to process conflicting communication styles can be independent of model size.

6 Test VLMs with OTOM$_{video}$

Now We test current VLMs' performance with OTOM$_{video}$, to guess the most appropriate *explanation* for given context with three types of NVCs: (1) visual cues (2) text (Socratic models) and (3) both. We test GPT-4o series, Qwen2.5-VL (Wang et al., 2024) series, and Ovis2 (Lu et al., 2024) series.

6.1 Methods

Visual Cues With our meticulously annotated frames that NVCs appear, the models receive visual token input in the form of a series of images,

representing the moments when NVC occurs during a dialogue.

Text Cues To test the capability without the visual recognition capability, we test VLMs with giving NVC name (e.g., Finger pointing).

Humans We only do *visual cue* setting because it is the least informative and the most realistic for the experiments. We ask 4 annotators to test OTOM$_{video}$ with identically sampled 50 questions, to calcuate accordance score.

6.2 Results

Model	Visual Cues	Text Cues	Both
Human	91.5	-	-
GPT-4o	73.6	73.9	79.0
GPT-4o-mini	65.3	79.0	77.1
Qwen2.5-VL-7B	58.3	74.2	73.2
Qwen2.5-VL-3B	51.0	57.6	59.9
Ovis2-8B	64.6	72.6	73.6
Ovis2-4B	49.0	56.7	55.4
Ovis2-2B	51.0	57.0	57.6
Ovis2-1B	30.6	28.0	32.8

Table 5: Performance of VLMs on OTOM$_{video}$. We test the models with only image frames of the moment of non-verbal cues shown (*Visual cues*), interpretation of NVC as text (*Text cues*, with just one black image), and both, along with the context of the dialogues.

Clear Human-AI Gap In Table 5, humans demonstrate superior performance (91.5% on visual cues) with high accordance (76%) compared to the best AI model (GPT-4o at 73.6-79.0%). This consistent gap across all settings (Visual Cues, Socratic Models, Both) suggests a fundamental limitation in current AI systems' ability to interpret contextual cues. Interestingly, performance in the 'Both' condition isn't necessarily better than individual conditions. This suggests that models don't always effectively integrate information from multiple modalities.

Model Size Correlation There's a clear scaling pattern where larger models generally perform better. For example, GPT-4o (73.6%) consistently outperforms GPT-4o-mini (65.3%), and Qwen2.5-VL-7B (58.3%) outperforms Qwen2.5-VL-3B (51.0%) with GPT-based models generally performing better.

7 Related Works

Theory of Mind Benchmarks to test ToM Capability in AI models have been developed a lot (Le et al., 2019; Kim et al., 2023; Jin et al., 2024), mainly focusing on false-belief task with text modality. MMToM-QA (Jin et al., 2024) proposes visual and contextual cue in household simulator, still focusing on asymmetric information and exclude dynamic human motion interpretation. Ma et al. (2023) develop multi-agent interaction and test ToM in a simple grid world, but does not fully capture natural human gesture or body language.

NVC Understanding NVC datasets are built in video understanding domain to classify the appropriate emotion state of the character in the video (Luo et al., 2020; Liu et al., 2021; Huang et al., 2021; Wicke, 2024). But they lack either (1) fine-grained emotion label (2-28, only focused on emotions), (2) action labels (detailed time line and which action the character is behaving). We develop OTOM$_{video}$, mapping the fine motion in the movie clip and its interpretation in the inner mind.

8 Conclusion

We present OBSERVABLETOM, a comprehensive framework for evaluating the understanding nonverbal social cues and their context-dependent meanings. Through extensive empirical analysis using both text and video datasets, we have identified significant gaps between current AI systems and human performance. These results highlight the need for fundamental advances in understanding NVCs that integrate contextual information and handle ambiguity.

9 Limitations

Limited 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.

10 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

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.

Ray L. Birdwhistell. 1970. *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press, Philadelphia, PA.

Su Lin Blodgett, Hal Daumé III, Michael Madaio, Ani Nenkova, Brendan O'Connor, Hanna Wallach, and Qian Yang, editors. 2022. *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2019. pyannote.audio: neural building blocks for speaker diarization. *Preprint*, arXiv:1911.01255.

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.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, 2(1).

Paul Ekman, Maureen O'Sullivan, Wallace V Friesen, and Klaus R Scherer. 1991. Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior*, 15(2):125–135.

Don Samitha Elvitigala, Denys JC Matthies, and Suranga Nanayakkara. 2020. Stressfoot: Uncovering the potential of the foot for acute stress sensing in sitting posture. *Sensors*, 20(10):2882.

Thompson Ewata. 2016. Meaning and nonverbal communication in films. *Issues in Language Linguistic: Perspectives from Nigeria*, 3.

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.

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

Alhakam Hameed Ali Hameed and Bushra Ni'ma Rashed. 2018. The social interaction of language in a comic series: A sociolinguistic study. *International Journal of Research in Social Sciences and Humanities*, 8(4):238–251.

Jinni A Harrigan. 2005. Proxemics, kinesics, and gaze. *The new handbook of methods in nonverbal behavior research*, pages 137–198.

Maria Hartwig and Charles Bond. 2014. Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28.

Agata Hołobut and Jan Rybicki. 2020. The stylometry of film dialogue: Pros and pitfalls. *Digital Humanities Quarterly*, 14(4).

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.

Dell Hymes. 1974. Ways of speaking. *Duranti, Alessandro. Linguistics Anthropology. A Reader. Oxford, Blackwell Publishing*, pages 158–171.

Dell Hymes. 2013. *Foundations in sociolinguistics: An ethnographic approach*. Routledge.

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

joblo. 2025. Joblo movie clips.

Kheryadi and Afif Suaidi. 2022. Digital ethnography of students' communication on whatsapp: An empirical study of native bantenese. *Journal of Linguistics and English Teaching Studies*, 7(1):74–82.

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*.

Michael Kipp and Jean-Claude Martin. 2009. Gesture and emotion: Can basic gestural form features discriminate emotions? In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–8. IEEE.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. 1978. *Nonverbal communication in human interaction*, volume 1. Holt, Rinehart and Winston New York.

Mina Kovačić. 2024. *Nonverbal Communication in Interrogation: Deception and Decoding Signals*. Ph.D. thesis, Josip Juraj Strossmayer University of Osijek. Faculty of Humanities and ....

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.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152.

lionsgate. 2025. Lionsgate movies.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10631–10642.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *Preprint*, arXiv:2405.20797.

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*.

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*.

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.

Esohe Mercy Omoregbe and Osasogie Christa Idada. 2020. Language use in social media and natural language. *International Journal of Humanities and Social Science*, 10(2):1–10.

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.

Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2023. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. *arXiv preprint arXiv:2309.15656*.

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.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zoya Shafique, Haiyan Wang, and Yingli Tian. 2023. Nonverbal communication cue recognition: A pathway to more accessible communication. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5666–5674.

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.

Sandra Tõnts. 2019. Chatbots: Will they ever be ready? pragmatic shortcomings in communication with chatbots. Master's thesis, Politecnico di Milano. MSc Thesis in Digital and Interaction Design.

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.

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.

Mark Yager, Beret Strong, Linda Roan, David Matsumoto, and Kimberly Metcalf. 2009. Nonverbal communication in the contemporary operating environment. page 91.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Miron Zuckerman, BM De Paulo, and R Rosenthal. 1981. Verbal and nonverbal.

A Human Evaluation Detail

A.1 Annotator Selection and Sampling

For our annotation process, we carefully select graduate students who, while not native English speakers, demonstrated high English proficiency sufficient for the task requirements. Considering the cognitive demands of the task and the importance of maintaining high-quality annotations, we implement a subsampling approach. We select 50 identical data points for all annotators to evaluate, enabling us to measure inter-annotator agreement. To ensure fair compensation, we establish a minimum hourly wage of $15 for their participation in the annotation process.

A.2 Data Preparation and Annotation Process

The authors conduct the primary data preparation and annotation process due to the complexity and precision required. This involves:

- Multiple review cycles (minimum 6 viewings) of each video to ensure accurate NVC identification
- Manual correction of STT (Speech-to-Text) outputs
- Refinement of speaker separation Direct annotation of NVC instances and associated emotional states
- Frame extraction based on STT timestamps with manual verification and correction

A.3 Interface

We utilize Label-studio[2] to create an intuitive labeling interface. The interface design is illustrated in Figure 6.

B 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)
- Qwen2.5-32B-Instruct (Yang et al., 2024)
- Qwen2.5-14B-Instruct (Yang et al., 2024)
- Qwen2.5-7B-Instruct (Yang et al., 2024)

[2]https://labelstud.io/

- Qwen2.5-3B-Instruct (Yang et al., 2024)
- Qwen2.5-1.5B-Instruct (Yang et al., 2024)
- Qwen2.5-VL-7B-Instruct (Wang et al., 2024)
- Qwen2.5-VL-3B-Instruct (Wang et al., 2024)
- Ovis2-8B (Lu et al., 2024)
- Ovis2-4B (Lu et al., 2024)
- Ovis2-2B (Lu et al., 2024)
- Ovis2-1B (Lu et al., 2024)

C Data Source

C.1 SPEAKING model

The SPEAKING model, developed by Dell Hymes as part of his ethnography of speaking methodology, provides a systematic framework for analyzing components of linguistic interaction (Hymes, 1974). This sociolinguistic model, represented through the mnemonic S-P-E-A-K-I-N-G, encompasses eight critical divisions: Setting and scene, Participants, Ends, Act sequence, Key, Instrumentalities, Norms, and Genre, which collectively enable researchers to examine the contextual elements essential for linguistic competence.

It is verified that the SPEAKING framework is versatile enough to be applied to analyze any event of communication, even those mediated by modern technology. In a study of group chats and messaging with WhatsApp (Kheryadi and Suaidi, 2022), researchers found that casual digital discourse follows systematical patterns, adjusting factors in the SPEAKING framework. In the studies on social media (Omoregbe and Idada, 2020; Hameed and Rashed, 2018), the framework is utilized to analyze special discourse in SNS including tagging someone and using hashtags. Furthermore, SPEAKING supports current studies on AI communication. In a study of chatbot interaction (Tõnts, 2019), category of user feedback is quantitatively identified with the SPEAKING framework. Researchers of voice-based agent design more context-aware AI dialogues with factors in the framework (Blodgett et al., 2022).

1. **Setting and Scene**: Refers to the physical circumstances (time and place) and psychological setting of the speech act. For example, a university lecture hall (setting) during a formal academic presentation (scene).
2. **Participants**: Includes both speakers and audience members involved in the communication. This encompasses direct addressees and indirect listeners, such as students actively participating in a class discussion and those passively listening.
3. **Ends**: Represents the purposes, goals, and expected outcomes of the speech event. For instance, in a job interview, the interviewer's end might be to assess candidate qualifications, while the interviewee's end is to secure employment.
4. **Act Sequence**: Describes the order and organization of the speech event. For example, in a formal debate, this includes opening statements, rebuttals, cross-examination, and closing arguments in a specific sequence.
5. **Key**: Indicates the tone, manner, or spirit of the speech act. This might range from serious (as in a funeral eulogy) to playful (as in casual conversation among friends).
6. **Instrumentalities**: Encompasses the channel, forms, and styles of speech used. This includes whether communication is verbal or written, formal or informal, and the choice of language or dialect. For example, using formal academic language in a conference presentation.
7. **Norms**: Refers to the social rules governing interaction and interpretation within the speech event. For instance, turn-taking protocols in business meetings or expectations about interruptions in casual conversations.
8. **Genre**: Identifies the type of speech event or act, such as lectures, sermons, casual conversations, or formal speeches. Each genre has its own set of expected patterns and conventions.

C.2 The dictionary of body language

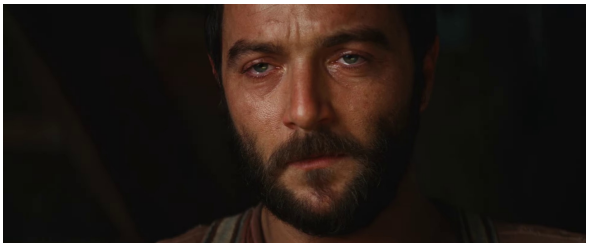
Joe Navarro is a behavioral analysis expert who served in the FBI for over 25 years, and his book Navarro (2018) is widely cited in psychology and rhetoric. Based on his extensive experience, he compiles 407 reliable NVCs (nonverbal cues). Navarro's key assertion that 'you must analyze behavioral clusters rather than single behaviors' has been emphasized by researchers including Paul Ekman, Mark Frank, and Aldert

Vrij (Ekman et al., 1991; Hartwig and Bond, 2014). 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 observations, such as foot direction indicating anxiety (Elvitigala et al., 2020) and lip-biting patterns (Kovačić, 2024), have been validated by subsequent research in nonverbal communication.

C.3 Movie Clips

The use of movie clips as data sources for NVC analysis is tested in several investigations. Film scenes provide controlled yet naturalistic environments for observing human nonverbal behavior, as evidenced by widely-used datasets like AFEW (Dhall et al., 2011) and CAER (Lee et al., 2019). While acknowledging that scripted interactions differ from spontaneous ones, research has shown that well-acted scenes effectively capture essential nonverbal human behaviors (Hołobut and Rybicki, 2020). The multimodal nature of film data - synchronizing facial expressions, gestures, and vocal intonations - makes it valuable for comprehensive NVC analysis (Ewata, 2016). Comparative analyses between scripted and spontaneous interactions have identified specific differences (Pilán et al., 2023), allowing researchers to account for these variations in their methodologies while leveraging the advantages of multimodal data available through film sources.

D Example of Model Answer



Dialogue

S0: You're sheltering enemies of the state, are you not?
S1: Yes.
S0: You're sheltering them underneath your floorboards, aren't you?
S1: Yes. [__]
S0: Point out to me the areas where they're hiding.
S1: (Pointing with finger)

Options

0: Feeling comfortable or attracted to someone
1: To emphasize or affirm what is being said
2: Internal conflict
3: Trying to absorb more light

Groundtruth: 2 ***Answer:*** 1

Table 6: Example of an incorrect output by GPT-4o, illustrating its inability to comprehend multimodal data in a human-like manner.

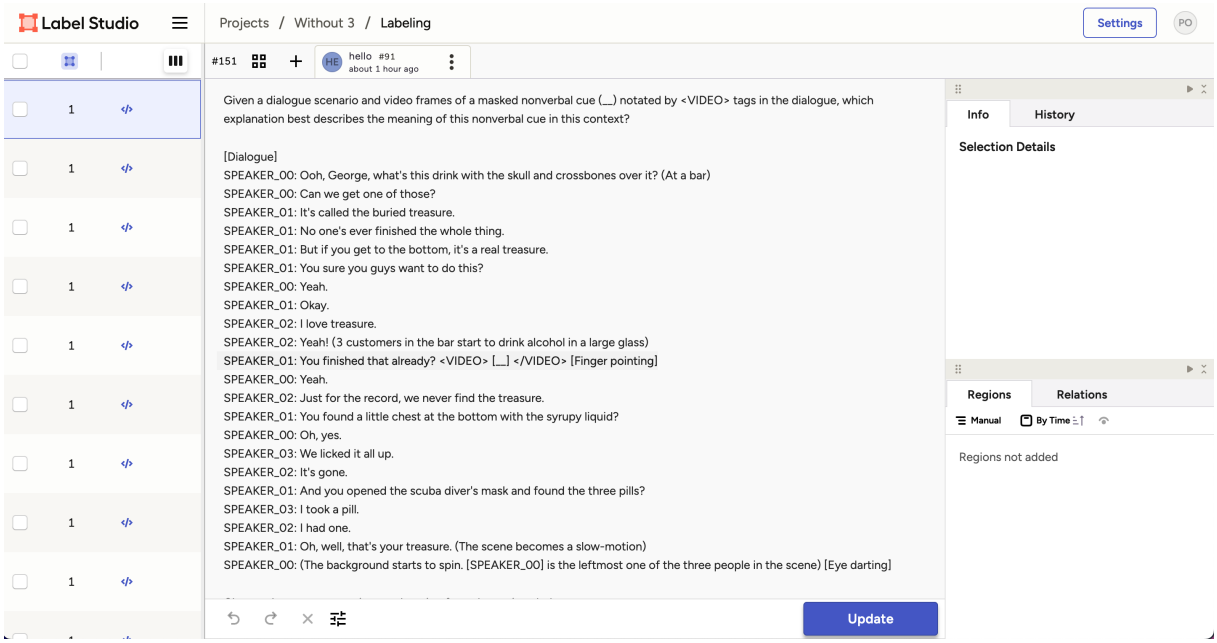


Figure 6: A labeling interface we use for §6.

Variables: explanation, options
Given the explanation of a nonverbal cue, please provide a plausible nonverbal cue from the options. [Explanation]: `{explanation}` [Options]: `{options}` Please answer with only the number (0-3).

Variables: nonverbal_cue, options
Given a nonverbal cue, please choose the most plausible explanation from the options. [Nonverbal Cue]: `{nonverbal_cue}` [Options]: `{options}` Please answer with only the number (0-3).

Variables: nonverbal_cue, explanation
Given a nonverbal cue and explanation, please determine if the explanation is valid. [Nonverbal Cue]: `{nonverbal_cue}` [Explanation]: `{explanation}` Please answer with only the label (valid/invalid).

Table 7: Prompts to test LLMs in §3.

Variables: dialogue, options, max_option, video_frames

Given a dialogue scenario and video frames of a nonverbal cue notated by <VIDEO> tags in the dialogue, which explanation best describes the meaning of this nonverbal cue in this context?
 	
[Dialogue]
`{dialogue}`
 	
Choose the most appropriate explanation from the options below:
`{options}`
 	
Answer with the number only (0-`{max_option}`).

Variables: dialogue, options, max_option, video_frames

Given a dialogue scenario and video frames of a masked nonverbal cue (__) notated by <VIDEO> tags in the dialogue, which explanation best describes the meaning of this nonverbal cue in this context?
 	
[Dialogue]
`{dialogue}`
 	
Choose the most appropriate explanation from the options below:
`{options}`
 	
Answer with the number only (0-`{max_option}`).

Table 8: Prompts for Dialogue-based Nonverbal Cue Understanding Tasks for VLMs in §6. The video_frames variable is a set of image frames that is directly input to model.

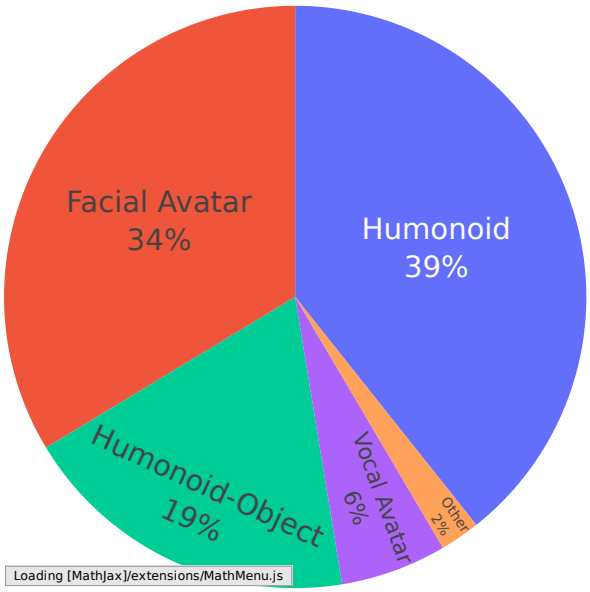


Figure 7: Body part categories in the data source (Navarro, 2018). For each nonverbal sign (e.g., Nose brushing), possible explanations are paired (e.g., Stress or discomfort, Questionable). We decompose the possible explanation into multiple elements.

Variables: mental_state, dialogue, options, max_option

Given the person's mental state and a conversation with a masked nonverbal cue (__), choose the most appropriate nonverbal cue that fits this mental state.

[Mental State]
{mental_state}

[Dialogue]
{dialogue}

Choose the most appropriate nonverbal cue from the options below:
{options}

Answer with the number only (0-{max_option}).

Variables: mental_state, dialogue, explanation, cue

Given the person's mental state and their nonverbal cue, determine if this explanation about the nonverbal cue is valid.

[Mental State]
{mental_state}

[Dialogue]
{dialogue}

Is this explanation [{explanation}] about the nonverbal cue [{cue}] valid based on this mental state? Answer with only 'valid' or 'invalid'.

Variables: mental_state, dialogue, cue, options, max_option

Given the person's mental state and their nonverbal cue, which explanation best describes the meaning of this nonverbal cue?

[Mental State]
{mental_state}

[Dialogue]
{dialogue}

[Nonverbal Cue]
{cue}

Choose the most appropriate explanation based on this mental state from the options below:
{options}

Answer with the number only (0-{max_option}).

Table 9: Prompts to test LLM performance in §5.

Variables: cue, interpretation, category

Analyze how the interpretation of nonverbal cues changes based on the SPEAKING model components. For each case:
Valid Context: Describe a situation where the nonverbal cue validly represents the given interpretation.
Invalid Contexts: Then describe how changing each SPEAKING component individually could invalidate this interpretation.

[Nonverbal Cue]: `{cue}`
[Possible Interpretation]: `{interpretation}` (`{category}`)

Valid Context:

Variables: situation_description, nonverbal_signs

Given nonverbal signs and situation descriptions, please generate a dialogue more than 10 lines.

RULES:

Use all the nonverbal signs ONLY ONCE, EXACTLY SAME WITH GIVEN in the dialogue. (e.g., [Head Nodding])
- Wrap the nonverbal sign in [BRACKETS] in the dialogue.
- Do not include the nonverbal sign in the utterance. Just space it in front of the dialogue or at the end of the dialogue.
- You can include some other nonverbal cues with (parentheses).
Please make the nonverbal cue to be subtle and natural.
- It's okay to include verbal cues contradicting the nonverbal sign (e.g., "I'm fine" [sad face]).
- DO NOT INCLUDE a direct metion of the nonverbal sign or its meaning in the dialogue, nor any other inner state of the character.
- The nonverbal sign should be a part of the dialogue, but not the main focus.
Try to stick to the given situation but you can change detail (relation between character, location, etc.) to make validity label more clear.
- If there are multiple settings in the situation description, you can choose or combine them as ONE which makes the explanations valid or invalid clearly.
- Please annotate informative character name or role in the dialogue. (e.g., Alice: "Hello, how are you?", Teacher: "I'm fine.")

[Situation Description]: `{situation_description}`
[Nonverbal Signs]: `{nonverbal_signs}`

[Step 1] Situation and Characters:

Variables: sign, explanation

Decompose the explanation of the given nonverbal sign with categorizing into one of these categories.

Disentangle a component into content (what it means), category (the mind state type), and details (probablistic situation of condition of that mind state if it exists).
If an explanation has multiple components, separate them with two newlines.
Content and details should be brief (e.g., "Stress", "Concern", "To show respect" "Communicate Religion") and should not contain the nonverbal sign itself.
If there is no specified detail, leave it as None.

[Categories]: Intention, Belief, Emotion, Knowledge, Desire, Percept
- Intention: A person's planned actions or goals they aim to achieve.
- Belief: What a person holds to be true about the world, whether accurate or not
- Emotion: A person's affective state or feelings in response to situations or stimuli
- Knowledge: Factual information or skills a person has acquired through experience or education
- Desire: What a person wants, wishes for, or hopes to obtain
- Percept: What a person is currently experiencing through their senses

[Nonverbal Sign]: `{sign}`
[Explanation]: `{explanation}`

OUTPUT:

Table 10: Prompts used to generate the dataset in §4.1.

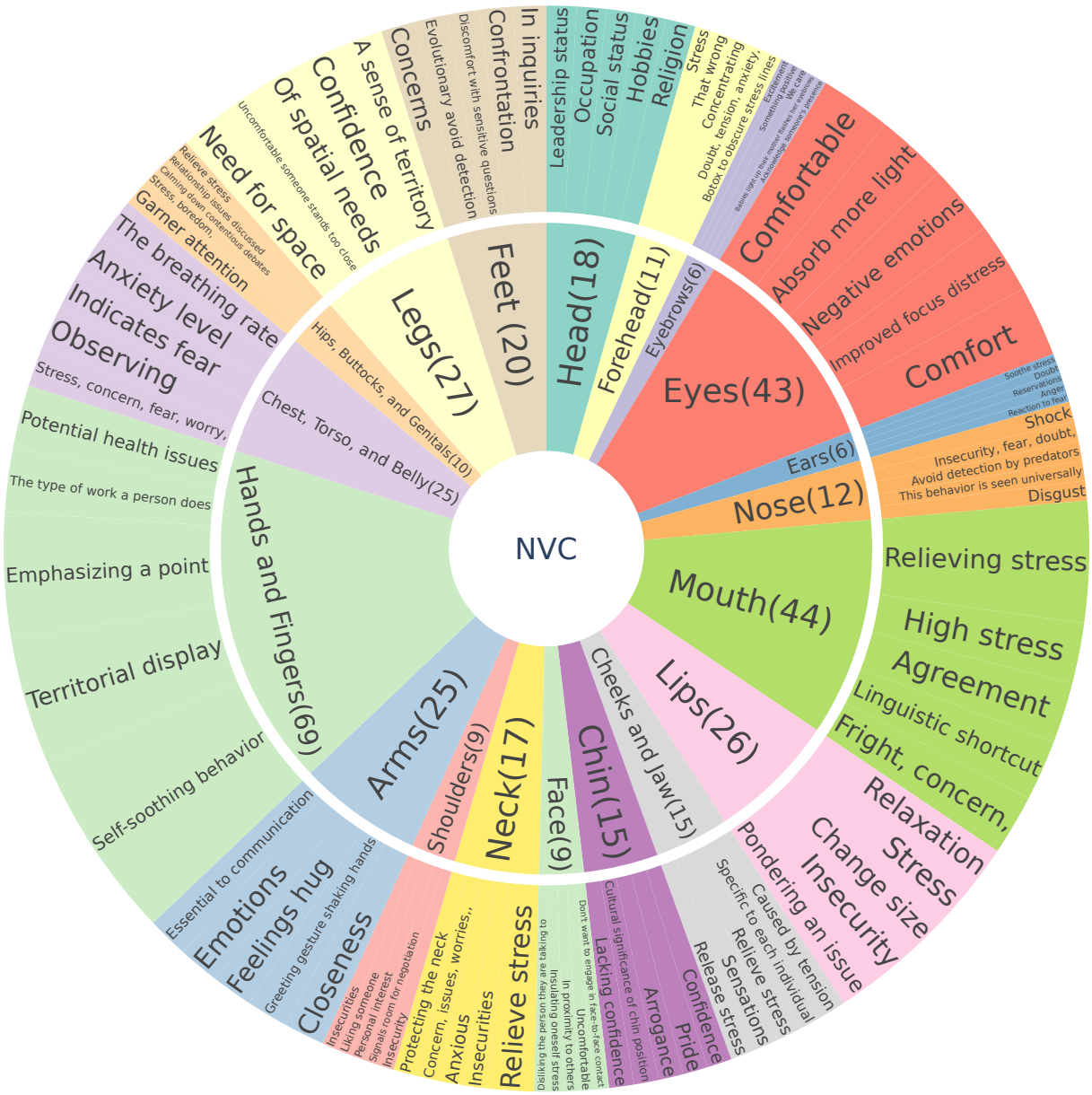


Figure 8: Anatomical categories in data source (Navarro, 2018) (see §2) and the five most frequent interpretations for each category. It includes 407 NVCs, and we get 1,114 interpretations by disentanglement (the prompt is provided in Table 10).

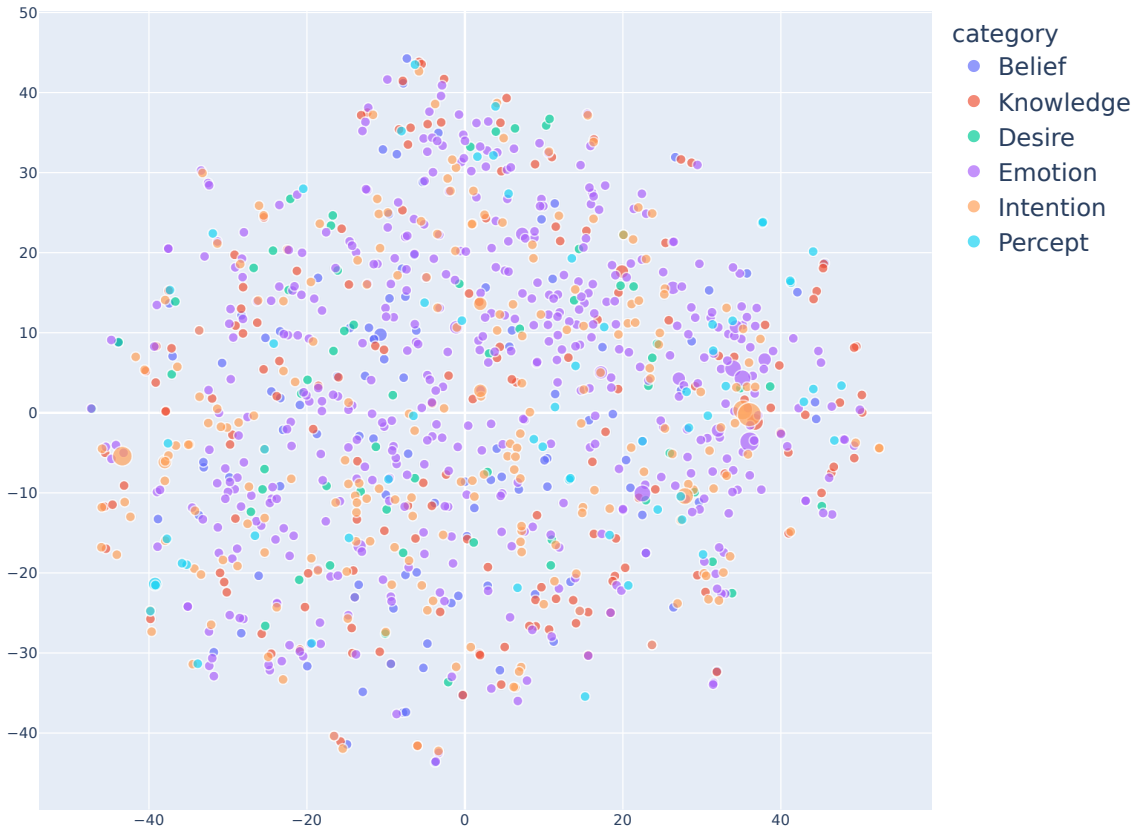


Figure 9: Scatter plot of semantic embeddings of explanations from our data source (Navarro, 2018). When we disentangle paragraphs into a list of multiple possibilities, we found that body language can convey more information about mental states beyond just emotions (43%) or beliefs (9%), including knowledge, intentions, desires, and perceptions.

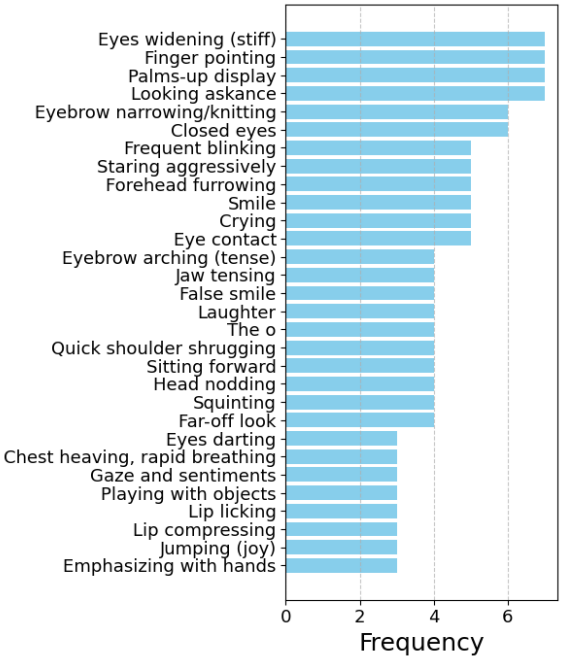


Figure 10: Top 30 frequency of NVCs in OTOM$_{video}$. Out of the 407 predefined NVCs, a total of 167 are utilized in the annotation process.

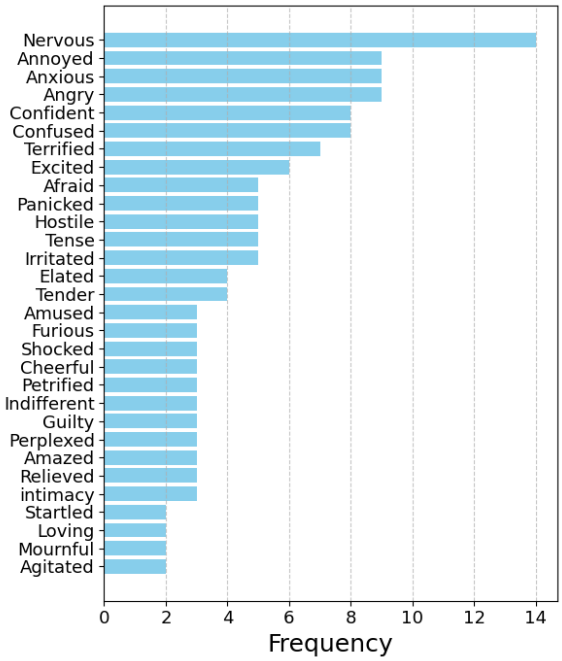


Figure 11: Top 30 frequency of explanations in OTOM$_{video}$. There are no restrictions on the formulation of explanations. A total of 185 are utilized in the annotation process.

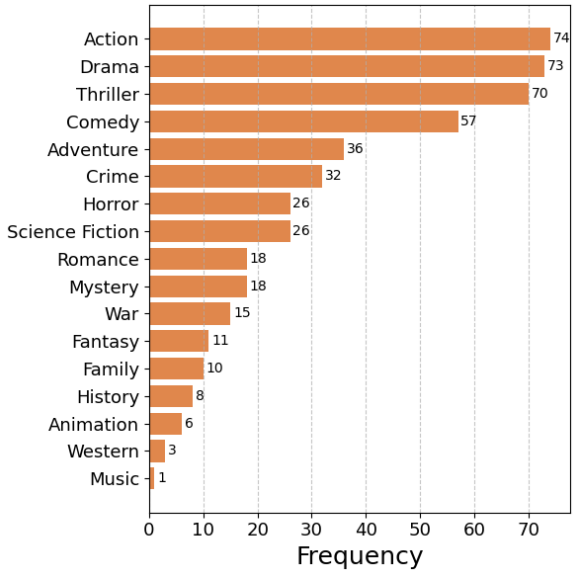


Figure 12: Distribution of genres of movie clips utilized in OTOM$_{video}$. Each video is allowed to include up to five genres, resulting in a total of 484 genre counts across 200 videos.

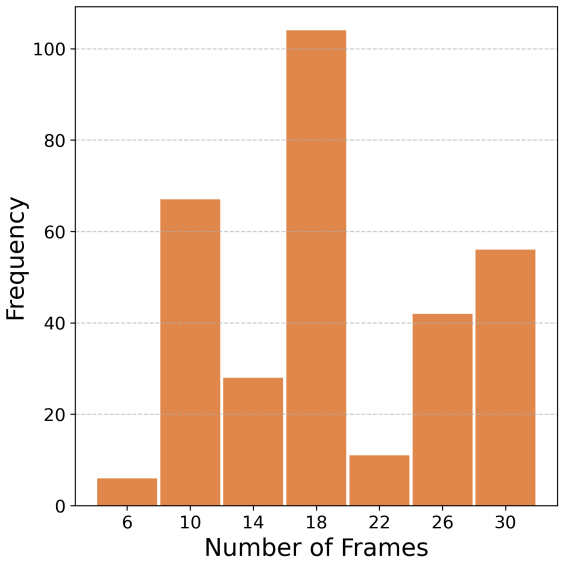


Figure 13: Distribution of the number of frames in OTOM$_{video}$. Given the conditions of 8 FPS and a 4-second duration, a maximum of 32 frames is allowed.