MIMEQA: Towards Socially-Intelligent Nonverbal Foundation Models

Hengzhi Li[1,2] **Megan Tjandrasuwita**[1] **Yi R. Fung**[1]
Armando Solar-Lezama[1] **Paul Pu Liang**[1]
[1]Massachusetts Institute of Technology [2]Imperial College London
Data: https://huggingface.co/datasets/hzli1202/MimeQA
Code: https://github.com/MIT-MI/MimeQA

Abstract

As AI becomes more closely integrated with peoples' daily activities, socially intelligent AI that can understand and interact seamlessly with humans in daily lives is increasingly important. However, current works in AI social reasoning all rely on language-only or language-dominant approaches to benchmark and training models, resulting in systems that are improving in verbal communication but struggle with nonverbal social understanding. To address this limitation, we tap into a novel data source rich in nonverbal social interactions – mime videos. Mimes refer to the art of expression through gesture and movement without spoken words, which presents unique challenges and opportunities in interpreting nonverbal social communication. We contribute a new dataset called MIMEQA, obtained by sourcing ~8 hours of videos clips from YouTube and developing a comprehensive video question-answering benchmark comprising 806 carefully annotated and verified question-answer pairs, designed to probe nonverbal social reasoning capabilities. Using MIMEQA, we evaluate state-of-the-art video large language models (VideoLLMs) and find that they achieve low accuracy, generally ranging from 20-30%, while humans score 86%. Our analysis reveals that VideoLLMs often fail to ground imagined objects and over-rely on the text prompt while ignoring subtle nonverbal interactions. We hope to inspire future work in AI models that embody true social intelligence capable of interpreting non-verbal human interactions.

1 Introduction

Social intelligence is integral to human interactions and enables nuanced understanding and communication with others [48, 28, 8]. There is increasing interest in developing socially intelligent AI systems that can understand and interact seamlessly with humans to help them in daily lives, such as stimulating empathic conversations in online mental health forums [67], assisting patients in geriatric care [27, 18], supporting children with autism spectrum conditions [32, 64], and helping educators in classroom teaching [78]. However, the majority of research towards socially intelligent AI focuses on language-only data and tasks (e.g., question-answering and dialogue) [39, 63], or multimodal data where language is often primary and nonverbal modalities (e.g., vocal and visual expression) are treated as second-class citizens [45, 85]. This results in a fundamental mismatch where today's foundation models are strong at language understanding but have a generally poor command of nonverbal social interactions; for example, nonverbal theory-of-mind [37], facial expression [30, 44], group social dynamics [68], and egocentric goal-oriented reasoning [34] are all challenges for today's language and multimodal foundation models.

To address these limitations, we tap into a novel data source rich in nonverbal social interactions – **mime performances**. Mimes refer to the art of expression through gesture and movement without spoken word [95], which presents unique challenges and opportunities for AI [60]. Since mime performances are devoid of speech, props, and actual objects, instead relying solely on the mime's ability to convey messages, emotions, and narratives through nonverbal communication, AI models

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

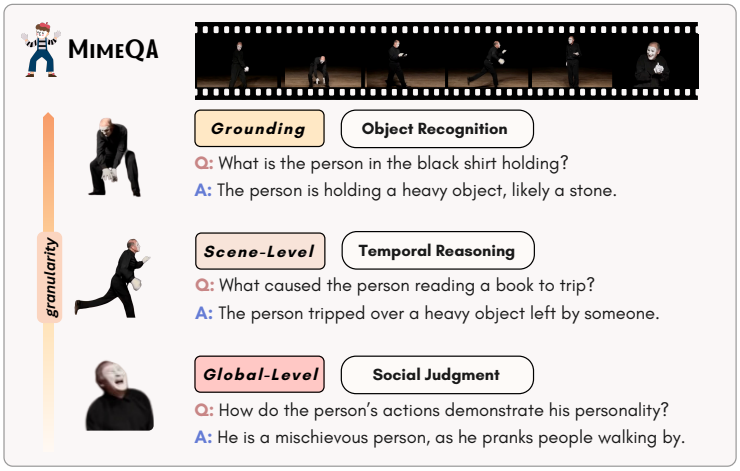


Figure 1: **MimeQA** is a new benchmark testing nonverbal social reasoning in multimodal large language models, with 101 videos of mimes (the art of expression through gesture without spoken words), and 806 question-answer pairs at three levels: 1) grounding the imagined object or activity, 2) scene-level understanding, and 3) global-level questions on holistic social comprehension. Most models achieve only 20-30% accuracy.

must have an acute understanding of human behavior, theory of mind, and the 'imagined' objects and actions they convey. Furthermore, mimes often depict complex interpersonal relationships and affective states from nonverbal interactions alone, without explicit narration and dialogue.

To systematically assess proficiency on these tasks, we create a benchmark called MimeQA, obtained by sourcing 221 mime videos spanning 8 hours of content from Youtube, annotating each video with questions ranging from local grounding tasks to broader theory of mind and social norm understanding, and meticulous verification of the annotations, resulting in 101 videos and 806 QA pairs. We benchmark state-of-the-art open-source and closed-source VideoLLMs and find that the overall accuracy ranges mostly from 20% to 30%, while humans perform 86%. Our extensive error analysis and ablations point to fundamental shortcomings of VideoLLMs' social and visual understanding capabilities, as common failure modes include failing to recognize imagined objects, misinterpreting nuanced social cues, and hallucinating responses based on the text input. We release our benchmark and evaluation framework at `https://github.com/MIT-MI/MimeQA` to drive future research toward verbal and nonverbal social intelligence in AI systems.

2 Theoretical Grounding & Related Work

Building **socially intelligent AI** involves creating agents that can sense, perceive, reason about, learn from, and respond to the affect, behavior, and cognition of other agents (human or artificial), and is a key part of AI as it becomes increasingly involved in our everyday lives [48]. To push the frontiers of socially intelligent AI, a rich body of work has examined various modalities, including language, video, audio, and more. For example, Gandhi et al. [22] evaluates the capabilities of AI to model human mental states from language to predict human goals and future behavior. Related work has also focused on extracting fine-grained visual features from gaze [70, 87], expressions [91, 92], and body language [81, 58, 84, 46]. Multimodal approaches have also been proposed to provide more holistic human intent understanding [45, 44, 49]. Wilf et al. [77] evaluate video understanding of social situation via question-answering, Jin et al. [35] evaluate theory of mind question answering on human household activities, and Li et al. [41] evaluate human intent understanding in videos.

Recent advances in **large multimodal models** have shown impressive video understanding capabilities in various domains, such as egocentric understanding and navigation [47], multimedia content analysis [42], and human language understanding [44, 74]. State-of-the-art enterprise models, such as Google Gemini [25] and GPT-4 [1], and open-source models such as Qwen-VL [61] and LLaVA-Video [88] have long context windows capable of handling video and audio inputs. These multimodal models have significantly improved performance on recent challenging video question-answering benchmarks [55, 47, 62, 19]. Despite significant progress, most existing models rely primarily on the language modality [44], resulting in commonsense biases in question prompts and, in extreme cases, good performance even without access to video at all [52]. Consequently, there is a lack of benchmarks that effectively evaluate the social intelligence of AI beyond language.

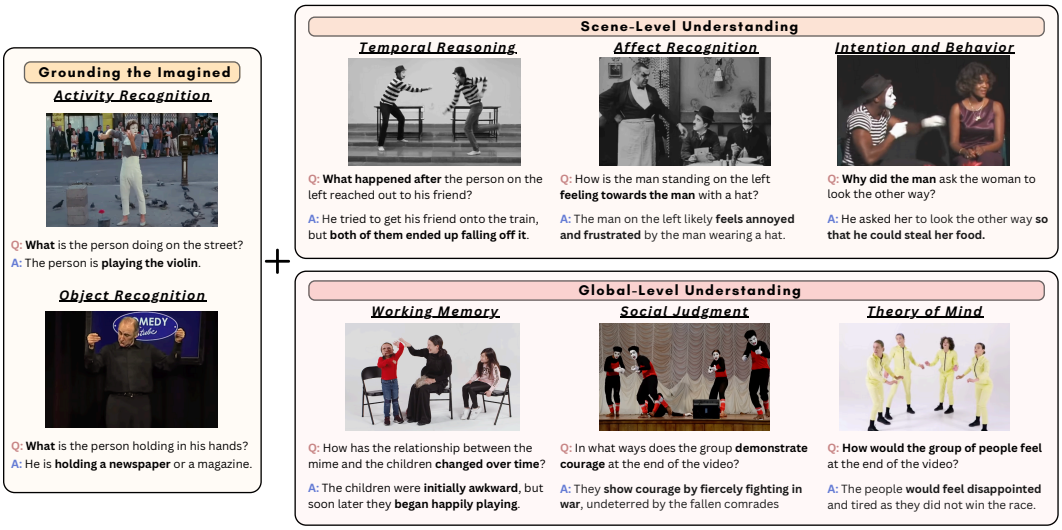


Figure 2: **Examples of MIMEQA question types. Left:** Grounding the imagined questions includes recognizing the activity or pretend object that the mime is acting out. **Top right:** Scene-level questions include temporal reasoning about a localized sequence of events, affect recognition questions about the emotional state of the characters, and intention and behavior questions that require interpreting the goals and motivations within a scene. **Bottom right:** Global-level questions involve working memory questions that probe understanding of the plot beyond localized sequences, social judgment questions about how the characters' actions adhere to cultural and social norms, and theory of mind questions about the characters' beliefs, desires, and motivation.

Mime performances serve as a good case for measuring nonverbal social intelligence. Mimes, or pantomimes when the performance has a coherent narrative, are often considered a peripheral form of communication due to their independence of speech and lack of structured conventions [50, 51]. Nevertheless, pantomimes have a crucial place in developing the human's natural language system; they are often seen as the fundamental building block to human language evolution, where systematic grammatical systems arise from increasingly complex gestural interaction over time [38, 53, 94, 16]. From a cognitive development perspective, Arbib [3, 4] posits that pantomimic gestures are crucial in the development from "language-ready" to "language use" brains, and studies have found that pantomime understanding is related to causal reasoning, working memory, and theory of mind capabilities [2, 24]. In human everyday communication, the highly iconic and transparent nature of pantomimic gestures leads to their frequent use in language-restrained settings, such as language impairment [17, 26], cross-cultural communication [57, 96], and neurodivergent communication [82]. Thus, mime performance presents a rich and untapped source for benchmark nonverbal social understanding in modern AI systems. While some prior works used mimes for evaluating action recognition [76, 11], to our best knowledge, MIMEQA is the first dataset to use mime performances to holistically evaluate multimodal foundation models' nonverbal social intelligence.

3 MIMEQA Dataset

We operationalize the opportunities and challenges of building nonverbal social intelligence through mime videos in a new open-ended video question-answering benchmark called MIMEQA. This benchmark consists of questions that evaluate social understanding at varying levels, from basic perception to complex reasoning about social dynamics across the full video.

3.1 Question Hierarchy

The MIMEQA questions are structured into three levels across the temporal scale, progressing from low-level visual recognition to scene-level interpretation and global-level cognitive reasoning. See Figure 2 for example questions for each category.

Grounding the Imagined. An important element of mime performances is its use of abstract iconic gestures or body movements to convey an imagined object or activity [95]. For example, a movement of flapping one's wings may represent a flying bird. These gestures are grounded in humans' embodied experience, and understanding their meaning is crucial for mimic communication [23, 94]. To measure the VideoLLMs' capabilities to ground these imagined objects and actions, our

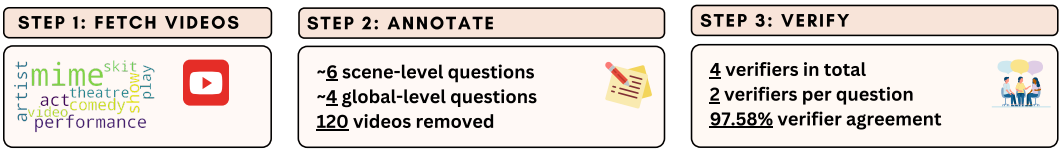


Figure 3: **Dataset construction pipeline:** 1) Collecting videos from YouTube with various search terms that are summarized by the word cloud. 2) Annotating approximately 6 grounding and scene-level questions and 4 global-level questions per video, removing 120 videos in the process. 3) Verifying the annotated questions and answers, with 97.58% verifier agreement.

first level of questions involves recognizing basic visual elements in the mime performance, such as objects and activities. This foundational perceptual information is a precursor for higher-level reasoning about interactions and intentions, as shown by Sibierska et al. [69].

Scene-Level. This level moves beyond perception to examine social interactions within a short video segment. Inspired by previous benchmarks [80, 77] and cognitive development research [9], we define three categories to assess fine-grained social understanding at the scene level.

- **Temporal reasoning** [73] requires structuring events into a causal chain linked by logical necessity and transitivity. This category involves identifying sequences of events in a scene and their temporal-causal relationships, beyond mere event ordering.
- **Affect recognition** [59] involves identifying and analyzing emotional states through nonverbal cues. Other than static emotion classification, this category also requires detecting subtle emotional shifts, group sentiment, and changes in expression.
- **Intention and behavior understanding** [7] involves inferring the motivations behind actions and interpreting how observed behavior reflects unobserved internal goals and mental states.

Global-Level. This level assesses the ability to synthesize and reason social information across multiple scenes. Unlike scene-level understanding, it prioritizes organizing and weighing social cues to form higher-order interpretations rather than isolated moments. Drawing from research on non-linguistic narrative comprehension [6, 40, 2], we define three categories to evaluate global social intelligence.

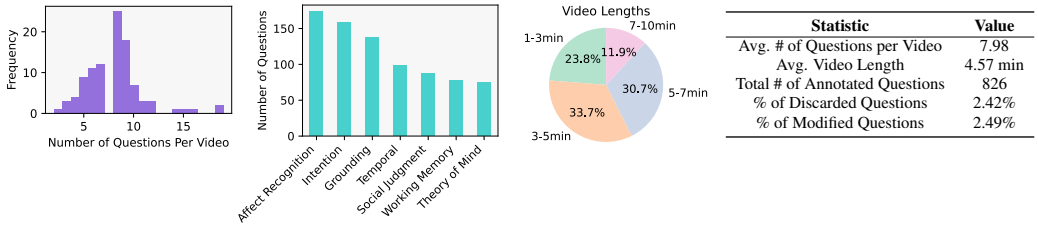
- **Working memory** [13] involves retrieving, integrating, and reasoning information across the entire video. Beyond single events, these questions require the ability to determine the relevance of past information, recall key events, and synthesize a coherent narrative.
- **Social judgment** [36] involves evaluating behaviors, assessing personality traits, and identifying social constructs like rapport, trust, and cooperation. This category requires comparing observations to social norms and counterfactual alternatives, highlighting unexpected or abnormal behavior.
- **Theory of mind** [5] measures the ability to infer beliefs, goals, and perspectives. This ability enables perspective-taking, reasoning about unseen motives, and anticipating how different individuals understand the same situation.

3.2 Dataset Construction

We summarize our dataset construction pipeline in Figure 3 and detail individual steps below.

Video collection. We collect videos from YouTube using various search terms that include the keyword "mime", downloading up to 50 videos per keyword. See Figure 3 for a word cloud of the search terms. We restrict video durations to between one and ten minutes. Additionally, we only select videos licensed under Creative Commons. This process yields a dataset of 221 videos.

Video validation and annotation. We asked two human annotators familiar with the question hierarchy to generate questions for each video, along with one-sentence answers to the question. The annotators are provided with a comprehensive description of the question hierarchy alongside a few examples per category. To ensure a diversity of categories, for each video, the annotators are asked to annotate approximately six scene-level questions, four global questions, and as many grounding questions as relevant, although the actual number of questions may vary based on the video. For grounding and scene-level questions, we asked them to provide start and end timestamps denoting the segment that the question is referring to. During the annotation process, annotators eliminated videos that lack a plot, are too difficult to understand, or explicitly involve language such as song lyrics or verbal explanations. We use the VGG Image Annotator [15] for all annotations.



Statistic	Value
Avg. # of Questions per Video	7.98
Avg. Video Length	4.57 min
Total # of Annotated Questions	826
% of Discarded Questions	2.42%
% of Modified Questions	2.49%

Figure 4: **MIMEQA dataset statistics.** Distribution of video lengths shows the range of short to long timescales. The distribution of the number of questions per video shows that each video is densely annotated, and the distribution of the number of questions per category is balanced.

Annotation verification. After an annotator has created a set of questions and answers for a video, a second person who has not seen the video verifies the quality of the annotation. The verifier is asked to watch the videos, answer the set of questions, and compare their answer with the originally annotated ground truth. The verifier then marks whether the two answers are consistent or otherwise provides suggestions to refine the questions. Finally, we manually review the verification results, remove any questions with inconsistent answers to avoid ambiguity, and refine the questions based on suggestions. By the end of this process, we reduced the original set of questions to 806 questions. See Figure 3 for an illustration of the dataset construction pipeline.

3.3 Dataset Statistics

Figure 4 contains MIMEQA's dataset statistics. The videos are densely annotated, with 806 total questions and most videos having more than five questions. MIMEQA has balanced questions across categories, with over 70 questions for each global category and over 100 questions for each local category. We provide additional analysis on MIMEQA's gender, cultural, and human behavior diversity in Appendix B.1.

4 Experiments

In this section, we evaluate closed and open-source VideoLLMs on the MIMEQA dataset. We detail the evaluation setup, present quantitative results, and conduct error analysis to understand model behavior in non-verbal social reasoning.

4.1 Experimental setup

We evaluate state-of-the-art closed- and open-source VideoLLMs on MIMEQA based on performance on video understanding benchmarks [19, 79]. For closed-source, we selected Gemini-1.5-Pro [25], Gemini-2.5-Pro [12], and GPT-4o [1], and we selected open-source models Qwen2.5-VL [61], LLaVA-Video [89], InternVL2.5 [10], and VideoLLaMA3 [86]. We use a standardized prompt, where we introduce the task of understanding mimes and subsequently ask a question, potentially including timestamps if it is a grounding or scene-level question. For models that do not natively support video input, we uniformly sample a number of frames with timestamps attached to the frames. See Appendix C.5 for the evaluation prompt template and Appendix C.2 for model settings.

To evaluate the model accuracy on our open-ended QA task, we use GPT-4o for LLM-as-a-judge [90] to automatically verify the model response against ground truth answers. A response is considered correct if it is semantically equivalent to the annotated ground truth. We evaluate the LLM grader quality on a sample of 352 questions and find that the automated grader aligns with a human grader 92.0% percent of the time. See Appendix C.5 for LLM grader prompt.

4.2 Results

We report the performance of open-source and closed-source models in Table 1. All models achieved low performance on the dataset: the open-source models achieve approximately 20% average accuracy, whereas GPT-4o achieves 31.3% and Gemini-2.5-Pro obtains 38.3%, significantly below the human baseline of 86.0%. This highlights the continued challenge for current models in visual abstraction and recognizing subtle social cues. In general, models perform better on global-level questions than on scene-level and grounding questions, suggesting that models struggle more with fine-grained video understanding compared to grasping the overall context of a video. Notably, models perform especially poorly on the grounding category, indicating a significant limitation in models' abstract visual cognition on imagined objects. GPT-4o, Gemini-1.5-Pro, and Gemini-2.5-Pro outperform open-source models by a factor of 2–3× across most categories.

Table 1: **Model accuracies on vision-language and language-only inputs across different questions.** VL=Video and text, L=Text only. **Avg**=Average overall performance. **GI**=Grounding the Imagined, **I**=Intention, **AR**=Affect Recognition, **T**=Temporal, **ToM**=Theory of Mind, **SJ**=Social Judgment, **WM**=Working Memory.

Model	Avg		Grounding		Scene-Level						Global-Level					
			GI		I		AR		T		ToM		SJ		WM	
	VL	L	VL	L	VL	L	VL	L	VL	L	VL	L	VL	L	VL	L
Qwen2.5-VL [61]	20.1	13.2	6.6	**9.5**	15.8	8.9	23.6	14.4	14.3	6.1	38.7	29.3	33.3	18.4	19.4	13.0
LLaVA-Video [89]	19.4	17.2	9.5	7.3	13.3	13.3	25.9	23.6	8.2	11.2	26.7	26.7	39.1	25.3	19.5	18.2
InternVL2.5 [10]	21.6	**20.6**	7.3	5.8	22.2	**17.1**	28.2	**25.9**	15.3	**12.2**	32.0	32.0	33.3	**43.7**	15.6	15.6
VideoLLaMA3 [86]	22.2	19.5	7.3	**9.5**	13.3	12.0	34.5	24.7	13.3	**12.2**	41.3	**36.0**	31.0	31.0	22.1	**20.8**
GPT-4o [1]	31.3	18.1	19.0	6.5	28.5	14.6	29.9	22.4	**30.6**	10.2	45.3	32.0	43.7	33.3	35.1	15.6
Gemini-1.5-Pro [25]	30.6	13.3	20.4	5.8	22.8	10.8	34.5	17.2	**30.6**	8.2	42.7	20.0	40.2	23.0	33.7	11.7
Gemini-2.5-Pro [12]	**38.3**	15.5	**28.4**	4.4	**31.6**	10.8	**43.7**	19.5	28.6	**12.2**	**54.7**	32.0	**51.7**	25.3	**39.0**	13.0
Human	86.0	-	89.8	-	87.3	-	83.9	-	88.8	-	93.3	-	80.5	-	76.6	-

To assess language bias in our dataset, we ablate the effect of video information by evaluating all models on text-only input, excluding video. We observe that models achieve higher accuracy on global-level questions than on scene-level ones without access to video. For example, without video, InternVL2.5 achieved 43.7% accuracy on the Social Judgment category, but only 5.8% on Grounding. This bias in global-level questions likely arises because some questions often include additional context to avoid referring to specific video segments, making it easier for models to infer information from annotations alone. Among open-source models, we observe an accuracy drop of 1-7% when transitioning from video to text-only evaluation. Interestingly, open-source models do not always benefit from video input. For example, both Qwen2.5-VL and VideoLLaMA3 observe a drop in accuracy in the Grounding category when provided with video, suggesting they struggle to integrate visual information effectively in question answering. In contrast, GPT-4o, Gemini-1.5-Pro, and Gemini-2.5-Pro demonstrate significantly better video comprehension, showing substantial accuracy improvements across all categories when provided with video input.

4.2.1 Improving Model Performance on MIMEQA

In this section, we attempt two distinct approaches to improve model performance on MIMEQA: supervised finetuning and the integration of explicit pose-based information. We also experimented with chain-of-thought prompting [75] in Appendix B.2, which yielded no gains in accuracy.

First, we evaluated whether supervised finetuning could enhance nonverbal social reasoning. From Table 2, we observe that finetuning the 72B Qwen2.5-VL model [61] on 80% of the MIMEQA benchmark improved overall accuracy from 22.5% to 26.6% on the held-out test set, reaching a level comparable to proprietary models like Gemini-1.5-Pro. The most significant gains occurred in categories requiring higher-level social reasoning, such as Intention (18.8% to 28.1%), Theory of Mind (44.4% to 55.6%), and Working Memory (23.5% to 47.1%). This suggests finetuning effectively improves the model's comprehension of nonverbal human behavior. However, Grounding performance remained low at 7.1%, indicating persistent challenges in inferring mimed objects and motivating our next experiment.

Table 2: **Fine-tuned vs. base Qwen2.5-VL on per-category accuracy.** FT = fine-tuned. GI = Grounding the Imagined, I = Intention, AR = Affect, T = Temporal, ToM = Theory of Mind, SJ = Social Judgment, WM = Working Memory. Finetuning Qwen2.5-VL on MIMEQA improves overall performance on test set.

Model	Avg	Grounding	Scene-Level			Global-Level		
		GI	I	AR	T	ToM	SJ	WM
Qwen2.5-VL	22.5	7.1	18.8	16.7	17.4	44.4	**47.4**	23.5
Qwen2.5-FT	**26.6**	**7.1**	**28.1**	**19.4**	**17.4**	**55.6**	31.6	**47.1**

To address the models' persistent low performance in grounding imagined objects and actions, we explored whether incorporating explicit human pose information as input could enhance performance on MIMEQA. We used PoseC3D [14], a skeleton-based action recognition model trained on the NTU RGB+D [66] dataset, to generate timestamped action labels for a 30% subset of MIMEQA at 10-second intervals. We retained only action predictions with confidence scores above 60%, which is chosen arbitrarily after manually checking a sample of the recognition results. These action labels were then included as additional input to the evaluation prompt of a Qwen2.5-VL-72B model. The prompt explicitly noted that the action recognition results are potentially unreliable.

Table 3: **Cross-dataset generalization results.** We report the average accuracy improvement from a five-fold cross-validation between MIMEQA, Social-IQ 2.0, and IntentQA. Models finetuned on MimeQA show good generalization to Social-IQ 2.0 and IntentQA, while models trained on other datasets struggle on MIMEQA.

	MimeQA Test	**Social-IQ Test**	**IntentQA Test**
Finetuned on MimeQA	**3.5% ± 3.1%**	**1.2% ± 3.0%**	2.6% ± 1.6%
Finetuned on Social-IQ 2.0 [77]	0.4% ± 2.3%	1.0% ± 2.3%	N/A
Finetuned on IntentQA [41]	1.1% ± 3.4%	N/A	**3.7% ± 1.2%**

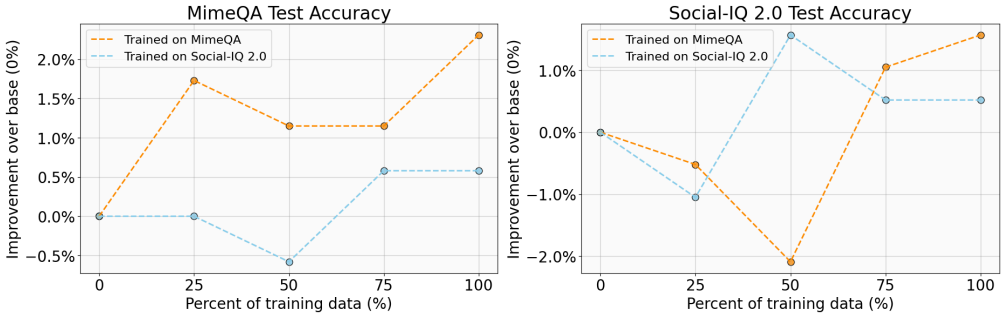


Figure 5: **Transfer analysis between MIMEQA and Social-IQ 2.0 [77].** Models fine-tuned on MIMEQA consistently generalize well to Social-IQ 2.0, while training on Social-IQ 2.0 yields little to no gains on MIMEQA. This highlights the distinct nonverbal social reasoning required in MIMEQA that is transferrable to other tasks.

This approach yielded a discernible trade-off. We saw a marginal decline in overall accuracy from 16.84% to 16.16% on this subset of MIMEQA. Notably, performance on fine-grained perceptual tasks improved, with accuracy on Grounding the Imagined increased from 2.33% to 6.98%, and Working Memory rose from 14.29% to 23.81%. Meanwhile, performance on higher-level social inference tasks dropped, with Theory of Mind decreasing from 52.17% to 34.78% and Social Judgment from 34.48% to 24.14%. We posit that the performance declines can be attributed to model hallucinations induced by inaccurate or overly literal action labels from the perception module, especially since PoseC3D has a limited number of 60 action labels. For example, a quarrel between actors is sometimes misclassified as "nausea or vomiting condition", which misled the model during reasoning.

Nevertheless, the improvement in the Grounding category represents a critical gain not achievable through supervised fine-tuning alone. This finding suggests that integrating explicit perception modules is a promising avenue for resolving fundamental grounding challenges in MIMEQA. However, future work must develop more sophisticated integration strategies that preserve low-level grounding benefits without compromising the model's capacity for high-level social reasoning.

4.2.2 Studying Transfer Between MIMEQA and Other Social Intelligence Tasks

To assess how well the nonverbal social reasoning skills learned from MIMEQAtransfer to other tasks, we performed a cross-dataset generalization experiment with two other social intelligence benchmarks: Social-IQ 2.0 [77] and IntentQA [41]. For each source dataset, we conducted five-fold cross-validation, finetuning a 7B Qwen2.5-VL model on four folds (80%) and validating on the remaining fold (20%) in each split. This allows us to measure both the finetuned model's in-domain performance (e.g., trained on MIMEQA, tested on MIMEQA) and its ability to generalize to the other two tasks. Since Social-IQ 2.0 and IntentQA are multiple-choice question-answering datasets, we adapt them to open-ended QA by using the correct choice texts as the target answers.

The results, summarized in Table 3, demonstrate the strong transferability of the skills learned from MIMEQA. On the Social-IQ 2.0 test, the model finetuned on MIMEQA achieves an average accuracy improvement of 1.2%, which is comparable to the 1.0% improvement from fine-tuning on Social-IQ 2.0 itself. Similarly, on the IntentQA test, the MIMEQA-finetuned model shows a 2.6% improvement, only slightly lower than the 3.7% gained from finetuning on IntentQA directly. These findings show that the nonverbal social understanding learned from MIMEQA effectively generalizes to broader social tasks. Conversely, models finetuned on Social-IQ 2.0 and IntentQA yielded much smaller accuracy improvements on MIMEQA at only 0.4% and 1.1%, respectively. Finetuning on MIMEQA offers significantly greater ($p < 0.05$) improvement of 3.5% on its own test set, highlighting the distinct nonverbal reasoning required by our benchmark.

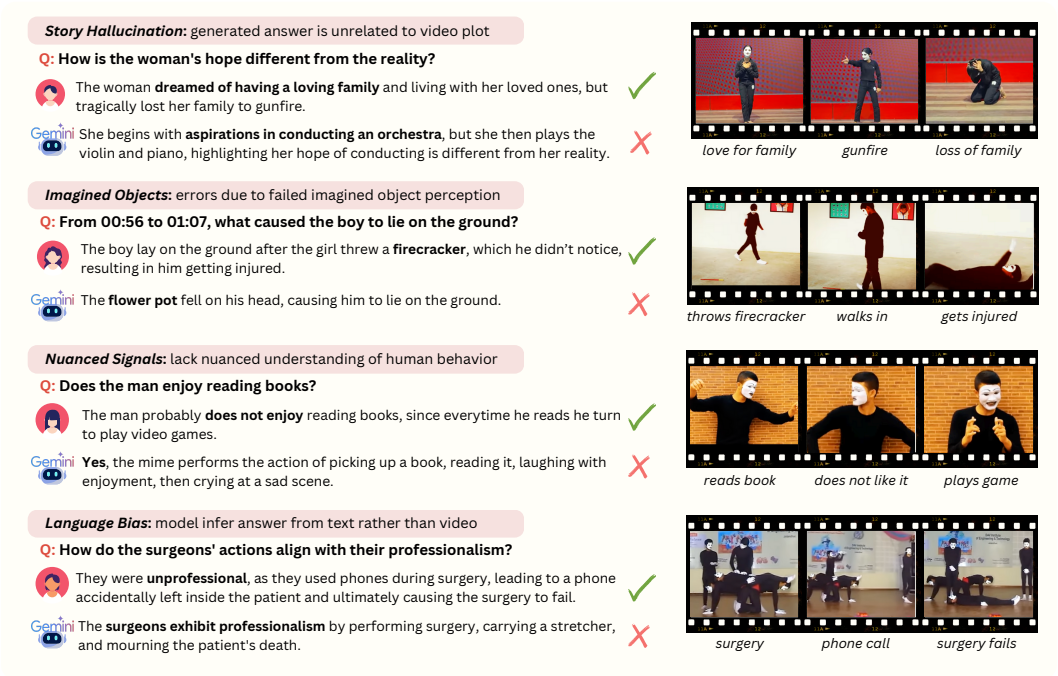


Figure 7: **MIMEQA model error examples.** We identify four common error categories. **Top-Bottom**. Story hallucination errors are when the model's response is unrelated to the video plot. Imagined objects denote errors where the model misidentifies the imagined objects. Nuanced signals denote instances of the model lacking a nuanced understanding of human behavior. Language bias denotes errors when the model is misled by the framing of the question and ignores the video.

To further investigate these performance differences, we progressively finetuned Qwen2.5-VL on increasing subsets (25%, 50%, 75%, 100%) of the MIMEQA and Social-IQ 2.0 train sets. As shown in Figure 5, increasing the training data from MIMEQA *consistently* improves accuracy on its own test set, while training on Social-IQ 2.0 yields negligible gains. We hypothesize this gap occurs because the verbal nature of common real-world videos in Social-IQ 2.0 means they often lack the expressive body language central to mime performances. In fact, finetuning on Social-IQ 2.0 did lead to improvements on Theory of Mind (21.1% to 27.6%) and Intention (8.8% to 11.7%) questions, particularly on questions with sufficient grounding context. However, Social-IQ finetuning degraded performance in categories requiring precise interpretation of mimed actions: Temporal accuracy dropped from 9.1% to 7.0%, and Working Memory fell from 17.0% to 14.2%. This drop was often caused by increased hallucination. For example, when asked "Why did the person on the left lose what he was holding?", the MIMEQA-tuned model correctly identified a quarrel, whereas the Social-IQ-tuned model misinterpreted the action as a dance routine. Thus, the social cues in Social-IQ 2.0 can be insufficient and even detrimental for tasks demanding robust nonverbal understanding.

Overall, our results demonstrate that finetuning on MIMEQA yields consistent and transferable gains across diverse social reasoning benchmarks, highlighting its unique contribution of social information not captured by existing datasets.

4.3 Error Analysis

We highlight the main sources of errors by the VideoLLMs on MIMEQA, focusing on Gemini-1.5-Pro. We plot the distribution of sources of errors in Figure 6.

Story hallucination from missing language grounding. One common pitfall pf today's best models on MIMEQA is hallucinating an answer that is plainly disconnected from the performance narrative. Due to the abstract and nonverbal nature of mime performances, VideoLLMs may

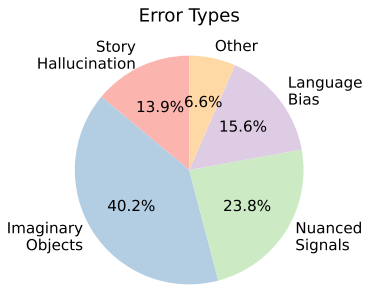


Figure 6: **Error types distribution.** We annotate the error types for 20 videos and plot the distribution.

interpret narratives in ways that deviate from commonsense. Figure 7 contains an example where the mime is acting as a woman who, initially living peacefully with family, tragically lost her family during a war. However, Gemini-1.5-Pro misunderstands the narrative and hallucinates that the mime is conducting an orchestra, which is completely unrelated from the video.

We hypothesize that the model hallucinations stem from the lack of language grounding in mime performances, which provide no verbal context as in existing social video datasets with spoken communication. To test this hypothesis, we examined how model performance varies between videos containing meaningful text—such as hand-held signs or banners indicating the topic of the performance—and those without text. We sample frames from videos at one frame per second and use EasyOCR [33] for text detection. A human then verifies the detected text, filtering meaningless texts like watermarks. We report model accuracy on videos with and without text in Table 4. We observe that most models, with the exception of LLaVA-Video, achieved higher accuracy on videos containing text, which highlights their dependence on language modality and explains their poor performance on MIMEQA.

Table 4: **Model performance on videos with and without text.** Text in the video frames is detected automatically with manual verification. All models except for LLaVA-Video have significantly improved performance on videos containing text.

Model	With Text	Without Text
Qwen2.5-VL [61]	**24.6**	15.5
LLaVA-Video [89]	19.2	**19.5**
InternVL2.5 [10]	**22.9**	20.3
VideoLLaMA3 [86]	**27.1**	17.3
GPT-4o [1]	**37.9**	24.5
Gemini-1.5-Pro [25]	**35.2**	26.0
Gemini-2.5-Pro [12]	**44.8**	31.8

Additionally, we investigate whether providing video titles as supplementary language context improves model accuracy. We select Qwen2.5-VL for experiment, and we report the results in Table 5, where we observe that incorporating titles in the input prompt enhances accuracy across most categories. These results highlight a fundamental limitation: models heavily rely on language cues for social commonsense reasoning. To advance nonverbal social intelligence, we must **rethink visual cognition** in multimodal foundation models, ensuring better alignment of social signals across diverse modalities, especially when verbal information is not present.

Table 5: **Model performance with and without video title provided.** T: text prompt includes title. **Avg**=Average performance across all questions. **GI**=Grounding the Imagined, **I**=Intention, **AR**=Affect Recognition, **T**=Temporal, **ToM**=Theory of Mind, **SJ**=Social Judgment, **WM**=Working Memory. Qwen2.5-VL's performance improves across almost all categories when given the title.

Model	Avg	Grounding	Scene-Level			Global-Level		
		GI	I	AR	T	ToM	SJ	WM
Qwen2.5-VL (with title)	**24.3**	**10.2**	**21.5**	21.8	**20.4**	**45.3**	**36.8**	**31.2**
Qwen2.5-VL (without title)	20.1	6.6	15.8	**23.6**	14.3	38.7	33.3	19.4

Failure to interpret imagined objects. Understanding mime performances requires the audience to imagine invisible objects or activities from fine-grained gestures and body language [69]. Our analysis suggests that models struggle to perceive imagined objects, leading to downstream reasoning errors in MIMEQA. For example, in Figure 7, a girl throws a firecracker on the ground, causing a boy to fall and appear injured. However, the model incorrectly identifies the firecracker as a flower pot. We also observe that the accuracy of Grounding is often positively correlated with correctness in other question categories (see Appendix B.3).

To assess the impact of misperceived imagined objects on reasoning accuracy, we qualitatively analyze sample questions and examine how model responses change as object references become more explicit. We provide a representative example in Figure 8. In this example, when initially asked what happens after the man in the video raises his hands, Gemini-1.5-Pro provides an incorrect response, misinterpreting the mime's action as holding a trapeze. However, when the question is augmented with a clear description of the imagined objects, which are two children the man lifts onto his shoulders, Gemini-1.5-Pro now correctly responds that the mime is juggling them in the air. Building upon prior studies critically examining foundation models' abstract visual cognition [29, 83, 65, 72], our findings highlight the need for **better human-AI perception alignment** [54] to advance multimodal social intelligence.

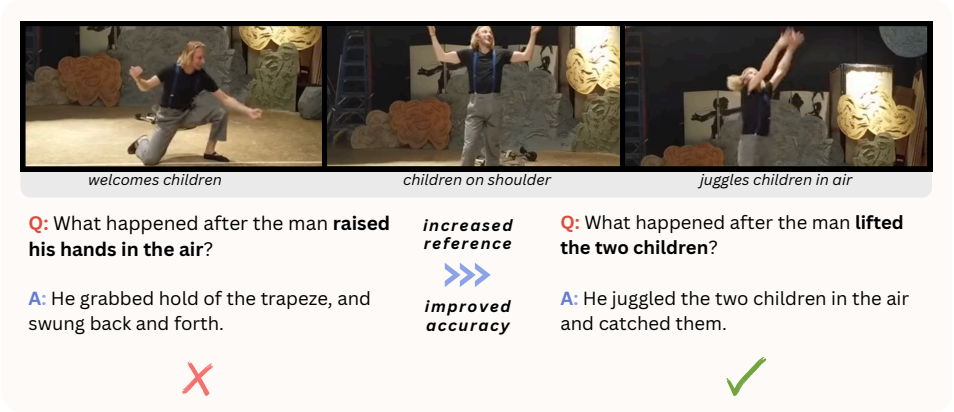


Figure 8: **Failure to interpret imagined objects impacts model's MIMEQA accuracy.** In this example, adding explicit reference to imagined objects in question allows Gemini-1.5-Pro to correctly answer the question.

Lack nuanced understanding of social signals. While models perform relatively well on Social Judgment and Theory of Mind compared to other categories, a closer examination reveals frequent errors stemming from a lack of nuanced understanding of human social signals. Figure 7 illustrates such a case: a man begins reading a book but eventually loses interest and switches to playing a game. When asked whether the man enjoys reading, Gemini-1.5-Pro incorrectly responds affirmatively, relying on a naive interpretation of his initial reading behavior rather than recognizing his loss of interest. These global-level questions require models to integrate various local signals into a comprehensive narrative, highlighting the limitations of VideoLLMs in the complexity of social reasoning. Therefore, there is a need for research on **fine-grained social reasoning** which has been relatively understudied [48, 49].

Language bias over video content. Finally, we observe that models often infer answers based on the question prompt rather than the video content, making them prone to biases in language inputs. For example, in Figure 7, the mimes depict a scene where surgeons use their phones during surgery, accidentally leaving one inside the patient, resulting in their death. However, Gemini-1.5-Pro blindly and incorrectly identifies the surgeons as acting professionally, relying on prior assumptions and biases from its language pretraining rather than interpreting the visual narrative. This analysis is further supported by models' text-only accuracy results in Table 1 which show that, particularly for open-source models, performance improves only marginally when video context is provided alongside the question text. This suggests that VideoLLMs' reliance on the question prompt rather than genuine video understanding.

The above findings underscore the need for multimodal models that effectively integrate all input modalities on language rather than over-rely on language, as also observed in other works [20]. Additionally, while social bias in language models has been widely studied [43, 21], our results emphasize the need to **understand and mitigate how these biases transfer** to multimodal social tasks, given the models' dependence on language.

5 Conclusion

Our MIMEQA benchmark highlights the crucial need for video LLMs to move beyond linguistic bias by integrating deeper non-verbal understanding for socially intelligent AI. By proposing mime understanding as a novel evaluation setting, we introduce a challenging yet valuable benchmark that requires models to interpret human gestures, emotional dynamics, and social interactions without explicit spoken dialogue. Our fine-tuning experiments reveal that while targeted training can improve higher-level social reasoning capabilities, fundamental challenges in grounding imagined objects persist. The asymmetric transfer patterns between MIMEQA and an existing social benchmark demonstrate that MIMEQA captures distinct aspects of social cognition, with mime-trained models generalizing well while models trained on conventional social datasets show minimal improvement on mime understanding. Our comprehensive analysis presents new research directions toward advancing the next generation of verbal and nonverbal socially intelligent foundation models.

Acknowledgement

MT is supported by the National Science Foundation (NSF) under Grant No. 2141064. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation. We thank the MIT Office of Research Computing and Data (ORCD) and the NVIDIA Academic Grant Program for GPU support.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Ines Adornetti, Alessandra Chiera, Valentina Deriu, Daniela Altavilla, and Francesco Ferretti. Comprehending stories in pantomime. a pilot study with typically developing children and its implications for the narrative origin of language. *Language & Communication*, 93:155–171, 2023.

[3] Michael Arbib. Toward the language-ready brain: biological evolution and primate comparisons. *Psychonomic bulletin & review*, 24:142–150, 2017.

[4] Michael Arbib. Pantomime within and beyond the evolution of language. In *Perspectives on Pantomime*, pages 16–57. John Benjamins, 2024.

[5] Janet Wilde Astington and Jennifer M Jenkins. Theory of mind development and social understanding. *Cognition & Emotion*, 9(2-3):151–165, 1995.

[6] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of developmental psychology*, 4(2):113–125, 1986.

[7] Sarah-Jayne Blakemore and Jean Decety. From the perception of action to the understanding of intention. *Nature reviews neuroscience*, 2(8):561–567, 2001.

[8] Cynthia Breazeal. Toward sociable robots. *Robotics and autonomous systems*, 42(3-4):167–175, 2003.

[9] Silas E Burris and Danielle D Brown. When all children comprehend: increasing the external validity of narrative comprehension development research. *Frontiers in Psychology*, 5:168, 2014.

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, January 2025.

[11] Hyundong Justin Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T. Chavez, and Jonathan May. Can vision language models understand mimed actions? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26744–26759, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1372. URL https://aclanthology.org/2025.findings-acl.1372/.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[13] Meredyth Daneman and Philip M Merikle. Working memory and language comprehension: A meta-analysis. *Psychonomic bulletin & review*, 3(4):422–433, 1996.

[14] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2969–2978, 2022.

[15] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535. URL https://doi.org/10.1145/3343031.3350535.

[16] Francesco Ferretti. On the influence of thought on language: a naturalistic framework for the pantomimic origins of human communication. *Frontiers in Psychology*, 14:1197968, 2023.

[17] Bibi Fex and Ann-Christin Månsson. The use of gestures as a compensatory strategy in adults with acquired aphasia compared to children with specific language impairment (sli). *Journal of neurolinguistics*, 11(1-2): 191–206, 1998.

[18] Kevin C Fleming, Jonathan M Evans, and Darryl S Chutka. Caregiver and clinician shortages in an aging nation. In *Mayo Clinic Proceedings*, volume 78, pages 1026–1040. Elsevier, 2003.

[19] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[20] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

[21] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[22] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding Social Reasoning in Language Models with Language Models, December 2023.

[23] Peter Gärdenfors. Demonstration and pantomime in the evolution of teaching. *Frontiers in psychology*, 8: 415, 2017.

[24] Peter Gärdenfors. The relations of demonstration and pantomime to causal reasoning and event cognition. In *Perspectives on Pantomime*, pages 58–77. John Benjamins, 2024.

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

[26] Susan Goldin-Meadow. *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. Psychology Press, 2005.

[27] Carina Soledad González-González, Verónica Violant-Holz, and Rosa Maria Gil-Iranzo. Social robots in hospitals: a systematic review. *Applied Sciences*, 11(13):5976, 2021.

[28] Hyowon Gweon, Judith Fan, and Been Kim. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023.

[29] Joy Hsu, Jiayuan Mao, Joshua B Tenenbaum, Noah D Goodman, and Jiajun Wu. What makes a maze look like a maze? *arXiv preprint arXiv:2409.08202*, 2024.

[30] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

[31] Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. Cultureclip: Empowering clip with cultural awareness through synthetic images and contextualized captions, 2025. URL https://arxiv.org/abs/2507.06210.

[32] Nikki Hurst, Caitlyn Clabaugh, Rachel Baynes, Jeff Cohn, Donna Mitroff, and Stefan Scherer. Social and emotional skills training with embodied moxie. *arXiv preprint arXiv:2004.12962*, 2020.

[33] Jaided AI. EasyOCR - ready-to-use optical character recognition with Pytorch. https://github.com/JaidedAI/EasyOCR, 2023. Accessed: 2025-02-15.

[34] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, December 2022.

[35] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.851. URL https://aclanthology.org/2024.acl-long.851/.

[36] Daniel Kahneman and Dale T Miller. Norm theory: Comparing reality to its alternatives. *Psychological review*, 93(2):136, 1986.

[37] Dora Kampis, Dóra Fogd, and Ágnes Melinda Kovács. Nonverbal components of Theory of Mind in typical and atypical development. *Infant Behavior and Development*, 48:54–62, August 2017. ISSN 0163-6383. doi: 10.1016/j.infbeh.2016.11.001.

[38] Adam Kendon. Reflections on the "gesture-first" hypothesis of language origins. *Psychonomic bulletin & review*, 24:163–170, 2017.

[39] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, 2023.

[40] Sanne JM Kuijper, Catharina A Hartman, Suzanne Bogaerds-Hazenberg, and Petra Hendriks. Narrative production in children with autism spectrum disorder (asd) and children with attention-deficit/hyperactivity disorder (adhd): Similarities and differences. *Journal of abnormal psychology*, 126(1):63, 2017.

[41] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974, 2023.

[42] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. VideoVista: A Versatile Benchmark for Video Understanding and Reasoning, June 2024.

[43] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[44] Paul Pu Liang, Akshay Goindani, Talha Chafekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Hemm: Holistic evaluation of multimodal foundation models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[45] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[46] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022.

[47] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[48] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in AI agents: Technical challenges and open questions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1143. URL https://aclanthology.org/2024.emnlp-main.1143/.

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

[50] David McNeill. Gesture and thought. In *Gesture and thought*. University of Chicago press, 2008.

[51] David McNeill. *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012.

[52] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024.

[53] Ana Mineiro, Patrícia Carmo, Cristina Caroça, Mara Moita, Sara Carvalho, João Paço, and Ahmed Zaky. Emerging linguistic features of sao tome and principe sign language, 2017.

[54] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.

[55] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024.

[56] OpenAI. OpenAI o3 and o4-mini System Card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, April 2025. Accessed: 2025-10-20.

[57] Gerardo Ortega and Asli Özyürek. Types of iconicity and combinatorial strategies distinguish semantic categories in silent gesture across cultures. *Language and Cognition*, 12(1):84–113, 2020.

[58] Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Bqa: Body language question answering dataset for video large language models, 2024. URL https://arxiv.org/abs/2410.13206.

[59] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[60] Deepika Phutela. The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9(4):43, 2015.

[61] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

[62] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

[63] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. 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 4463–4473, 2019.

[64] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294, 2012.

[65] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pages 1–11, 2025.

[66] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[67] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.

[68] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6163–6170, 2019.

[69] Marta Sibierska, Monika Boruta-Żywiczyńska, Przemysław Żywiczyński, and Sławomir Wacewicz. What's in a mime? an exploratory analysis of predictors of communicative success of pantomime. *Interaction Studies*, 23(2):289–321, 2022.

[70] Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence*, 284:103275, July 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2020.103275.

[71] Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[72] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers, 2025. URL https://arxiv.org/abs/2506.23918.

[73] Tom Trabasso, Paul Van den Broek, and So Young Suh. Logical necessity and transitivity of causal relations in stories. *Discourse processes*, 12(1):1–25, 1989.

[74] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[75] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[76] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021.

[77] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. https://github.com/abwilf/Social-IQ-2.0-Challenge, 2023.

[78] Hansol Woo, Gerald K LeTendre, Trang Pham-Shouse, and Yuhan Xiong. The use of social robots in classrooms: A review of field-based studies. *Educational Research Review*, 33:100388, 2021.

[79] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857, 2025.

[80] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[81] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. LLM knows body language, too: Translating speech voices into human gestures. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5004–5013, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.273. URL https://aclanthology.org/2024.acl-long.273.

[82] H Melis Yavuz, Bilge Selçuk, and Barış Korkmaz. Social competence in children with autism. *International Journal of Developmental Disabilities*, 65(1):10–19, 2019.

[83] Eunice Yiu, Maan Qraitem, Charlie Wong, Anisa Noor Majhi, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint arXiv:2407.17773*, 2024.

[84] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019.

[85] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[86] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding, January 2025.

[87] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human Gaze Assisted Artificial Intelligence: A Review. *IJCAI : proceedings of the conference*, 2020:4951–4958, July 2020. ISSN 1045-0823. doi: 10.24963/ijcai.2020/689.

[88] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[89] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning With Synthetic Data, October 2024.

[90] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[91] Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.861. URL https://aclanthology.org/2023.acl-long.861.

[92] Wenjie Zheng, Jianfei Yu, and Rui Xia. A unimodal valence-arousal driven contrastive learning framework for multimodal multi-label emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 622–631, 2024.

[93] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

[94] Jordan Zlatev, Przemysław Żywiczyński, and Sławomir Wacewicz. Pantomime as the original human-specific communicative system. *Journal of Language Evolution*, 5(2):156–174, 2020.

[95] Przemysław Żywiczyński, Sławomir Wacewicz, and Marta Sibierska. Defining pantomime for language evolution research. *Topoi*, 37:307–318, 2018.

[96] Przemysław Żywiczyński, Sławomir Wacewicz, and Casey Lister. Pantomimic fossils in modern human communication. *Philosophical Transactions of the Royal Society B*, 376(1824):20200204, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our core contribution is our benchmark of mime performance videos, which we use to evaluate the nonverbal social understanding of foundation models. Our claim that large vision language models fall short of human-like nonverbal understanding is substantiated by experiments in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not have theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details on the multimodal foundation models and hyperparameters that we used for inference and supervised finetuning are in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have released our data and code at https://github.com/MIT-MI/MimeQA.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiment settings for inference and supervised finetuning are detailed in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For experiments involving open-source models, we run multiple seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details on the compute resources are in Appendix C.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the code of ethics and discussed any potential harms of our work in Appendix A.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix A for a discussion of potential societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out

that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our dataset poses little risk as the dataset is sourced from YouTube videos under Creative Commons license.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 3, we explicitly mention that the dataset is sourced from YouTube videos under Creative Commons license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: In Section 3, we thoroughly document the question hierarchy.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We used human annotators to construct MIMEQA. The details are mentioned in Appendices A and D.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not conduct research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our work evaluates the performance of large vision-language models on our proposed dataset and describes experiments involving these models in Section 4 and additional details in Appendix C.4.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Limitations, Ethics Statement and Broader Impact

In this paper, we focused on benchmarking AI models beyond linguistic cues by exploring the domain of mimes – a uniquely rich form of nonverbal, embodied human communication. However, there are many other expressive, human-centered interaction modalities that remain underexplored and present promising avenues for future research, including interactive art installations, dance, cultural performances, and musical expressions. These domains offer distinct affordances for evaluating and enhancing AI's capacity for social understanding. Additionally, our current dataset is limited in size, and developing effective training methods that can achieve robust social reasoning from small data regimes remains an open challenge. Finally, the cultural and demographic contexts reflected in our benchmark primarily represent Western societies, and semantic subtleties vary across cultures [31]. A crucial direction for future work is to broaden the cultural scope of MIMEQA to encompass a more diverse range of social norms, behaviors, and modes of expression across global communities.

All human annotations and verifications, both for dataset construction and analysis, were conducted by the authors. Details of the annotation instructions can be found in Appendix D. All video data used in this article are publicly available under Creative Commons license, and none of the annotations contain personally identifiable information. MIMEQA is released under a CC-BY-NC-SA 4.0 license and is intended solely for research purposes.

Developing rich nonverbal social reasoning is crucial for AI systems to interact effectively with humans and enhance well-being. While our work aims to advance this capability, we acknowledge the potential risks associated with such advancements, including applications in mass surveillance that could infringe on individual privacy. We support efforts to mitigate potential misuse and to ensure that socially intelligent AI systems are developed and applied responsibly.

B Additional Results

B.1 Diversity Analysis of MIMEQA

We acknowledge that the interpretation of mime performances may be influenced by cultural context and perceived gender of the performer. To understand the cultural and demographic biases in MIMEQA, we randomly sampled 25% of the videos in the dataset and manually annotated two attributes using both the video content and its title/description: 1) whether the cultural context appeared Western or non-Western, and 2) the perceived binary genders (male/female) of the individuals depicted. Our analysis indicates that 56% of the sampled videos reflect non-Western cultural contexts, with many sources originating from Asian regions. In terms of gender, 88% of videos feature at least one male participant, and 68% include at least one female participant. A majority of videos contain both male and female participants. Thus, although MIMEQA was not explicitly curated to enforce cultural or gender diversity, these findings suggest a reasonably balanced representation across both dimensions, with no dominant cultural or gender skew.

We further elaborate on the diversity of nonverbal social cues in MIMEQA. We examined the reference answers for the "Grounding the Imagined" category and extracted all words ending in "-ing," yielding a list of action-related terms. After removing mime- or performance-specific words (e.g., "pretending," "portraying"), we obtained 68 unique action keywords, comparable to the range found in many action recognition datasets. We provide a word cloud of the action keywords in Figure 3. To further estimate the diversity of actions in MIMEQA, we ran PoseC3D [14], an action recognition model trained on the NTU RGB+D dataset [66], on 10-second clips from a 30% subset of MIMEQA videos. Out of 60 possible action categories, the model identified 46 distinct actions across the subset. While this is only a proxy measure, we believe it reflects the broad range of perceived nonverbal behavior captured in MIMEQA.

Figure 9: **Word cloud of actions in MIMEQA.** There are 68 distinct action keywords in MIMEQA's answers, indicating its human behavior diversity.

B.2 Does Chain-of-Thought Improve Accuracy on MIMEQA?

We additionally experiment whether model accuracy on MIMEQA could be improved with chain-of-thought (CoT) prompting [75]. To integrate CoT prompting, we add the following line to the end of evaluation prompt.

```
"Think step by step to answer the question.  Format the
final answer in a separate sentence like 'The answer is X'."
```

We report the results in Table 6. Overall, although CoT prompting slightly improves Qwen2.5-VL's accuracy on MIMEQA, for most models it slightly decreases performance. Taking a closer look, we find that, for example, GPT-4o with CoT prompting tends to hallucinate more for the Grounding the Imagined, Temporal Reasoning, and Affect Recognition categories, thereby leading to performance drop. This finding is in line with prior work [71] that shows CoT is not always helpful for tasks which lack symbolic reasoning.

Table 6: **Model performance on MIMEQA with and without chain-of-thought (CoT).** We find that CoT yields no significant accuracy improvement on MIMEQA.

Model	Without CoT	With CoT
Qwen2.5-VL	20.1%	**22.2%**
LLaVA-Video	**19.4%**	17.7%
InternVL2.5	**21.6%**	18.7%
VideoLLaMA3	**22.2%**	17.7%
GPT-4o	**31.3%**	29.5%

B.3 Correlation between Grounding and Other Question Categories

To analyze the effect of the model's inability to understand localized events, we compute the correlation between performance on Grounding the Imagined questions to the other scene-level and global-level questions. Intuitively, we would expect that a model's ability to perform grounding would correlate strongly with, for example, temporal understanding, as one needs to understand individual actions and objects before reasoning about a sequence of events.

In Table 7, we report the computed accuracy correlation for selected models. We find that Qwen2.5-VL's Grounding performance positively correlates with Temporal understanding. For LLaVA-Video, Grounding accuracy correlates with Affect Recognition and Theory of Mind. For Gemini-1.5-Pro, we see that Grounding performance contributes both to understanding localized temporal sequences as well as to a more holistic understanding of the video, as shown by higher correlation scores with Temporal understanding, Social Judgment, and Working Memory. For GPT-4o, Grounding performance correlates with Affect Recognition and Theory of Mind. See Figure 10 for correlation between all pairs of question categories when the input contains both video and language, and Figure 11 for all correlations when the input is language only. This suggests that improved understanding of the fine-grained visual cues would lead to a better grasp of the video plot, with the specific reasoning pathways that benefit from this grounding being model-dependent. Our results demonstrate that an important line of future work is to improve VideoLLMs' ability to reason without explicit objects or human-object interactions, which can bottleneck performance on holistic video understanding.

Table 7: **Performance correlation between grounding the imagined questions and other categories for selected models. I**=Intention, **AR**=Affect Recognition, **T**=Temporal, **ToM**=Theory of Mind, **SJ**=Social Judgment, **WM**=Working Memory.

Model	I	AR	T	ToM	SJ	WM
Qwen2.5-VL	-0.091	-0.013	0.330	0.003	0.030	0.132
LLaVA-Video	-0.040	0.329	-0.106	0.228	0.129	-0.050
GPT-4o	-0.122	0.179	-0.040	0.163	-0.014	-0.156
Gemini-1.5-Pro	0.146	-0.053	0.301	-0.088	0.313	0.302

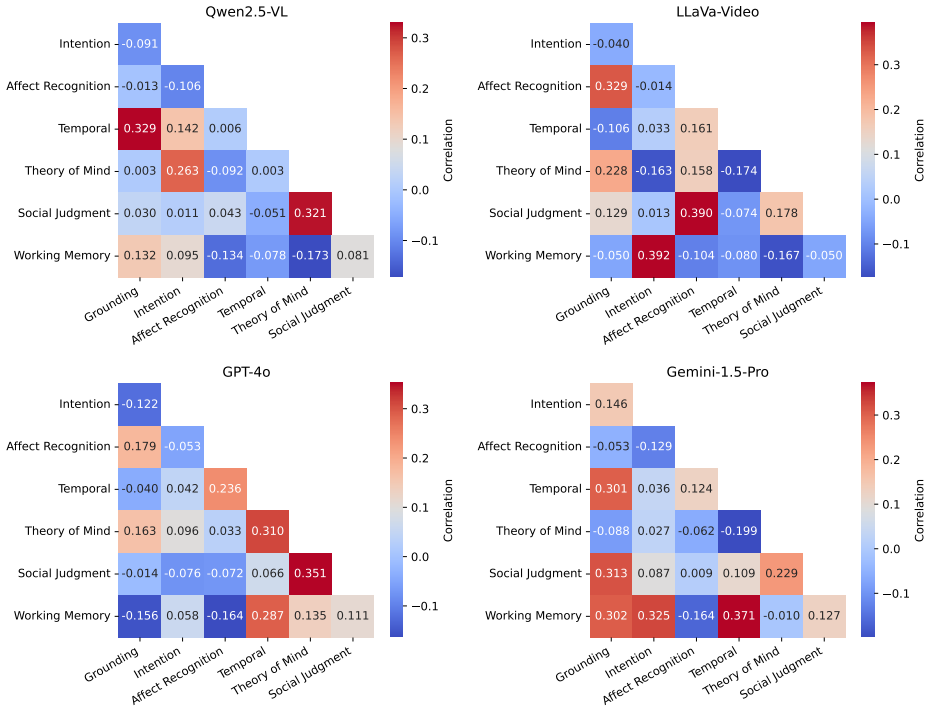


Figure 10: **Question type performance correlation matrices on video and text input.** For each video, we compute the accuracy over all question types and plot the correlation between accuracies on different question types. From left to right, we show the correlation matrices for Qwen2.5-VL, LLaVA-Video, Gemini-1.5-Pro, and GPT-4o.

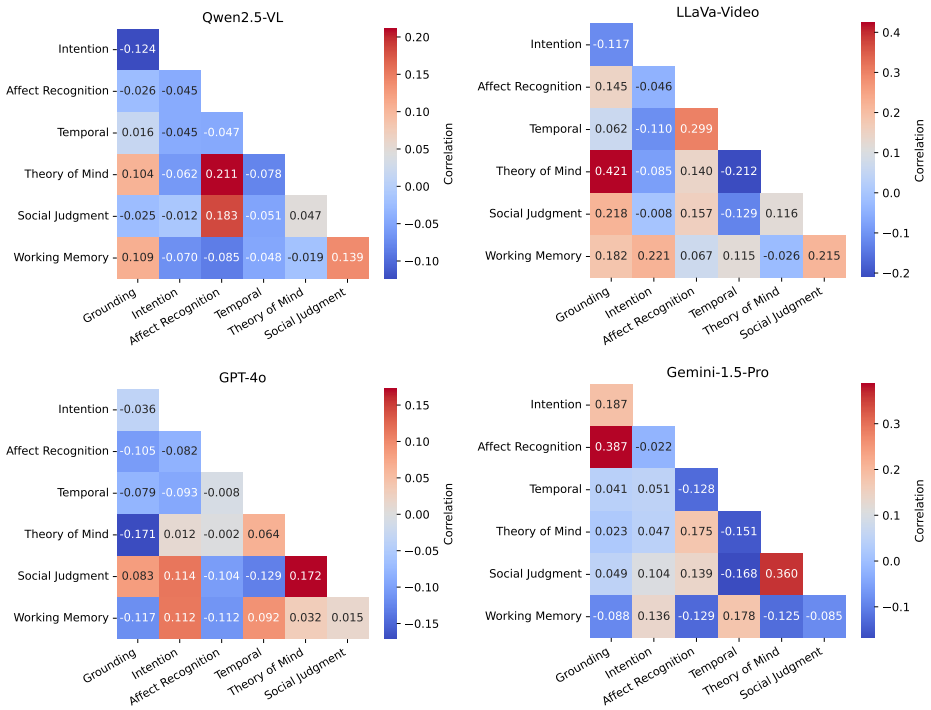


Figure 11: **Question type performance correlation matrices on only text input.** To ablate the effect of video, we perform inference for all models with only the text prompt. For each video, we compute the accuracy over all question types and plot the correlation between accuracies on different question types. From left to right, we show the correlation matrices for Qwen2.5-VL, LLaVA-Video, Gemini-1.5-Pro, and GPT-4o.

B.4 How Does Reasoning Vision-Language Models Perform on MIMEQA?

We evaluated the performance of frontier vision-language models not specifically fine-tuned for video understanding, focusing on OpenAI's GPT-o3 [56] on the MIMEQA benchmark. Following a similar setup to GPT-4o, videos were sampled at 1 frame per second to a maximum of 64 frames. The results, presented in Table 8, show that GPT-o3 (34.4) demonstrates exceptional strength in global-level questions requiring sophisticated reasoning. It significantly outperforms all other models on Theory of Mind (62.7) and Working Memory (41.6). It also performs well on scene-level Intention (32.3) and Temporal (34.7) reasoning.

However, these advanced reasoning capabilities appear to come at the cost of visual perception. GPT-o3 struggles with perception-heavy tasks, lagging behind Gemini-2.5-Pro in Affect Recognition (32.3 vs. 43.7). Notably, it scores just 13.1 on Grounding the Imagined, falling significantly behind both GPT-4o (19.0) and Gemini-1.5-Pro (28.4). This highlights a potential trade-off where language reasoning gains may compromise foundational visual grounding capabilities.

Table 8: **GPT-o3 accuracy on MIMEQA compared to selected models.** GPT-o3 shows the most significant gains on global level questions.

Model	Avg	Grounding	Scene-Level			Global-Level		
		GI	I	AR	T	ToM	SJ	WM
GPT-4o	31.3	19.0	28.5	29.9	30.6	45.3	43.7	35.1
Gemini-2.5-Pro	**38.3**	**28.4**	31.6	**43.7**	28.6	54.7	**51.7**	39.0
GPT-o3	34.4	13.1	**32.3**	32.2	**34.7**	**62.7**	44.8	**41.6**
Human	86.0	89.8	87.3	83.9	88.8	93.3	80.5	76.6

C Additional Experimental Details

C.1 Video Collection Details

We searched YouTube for videos using eight keywords: `mime performance`, `mime act`, `mime artist`, `mime comedy`, `mime theatre`, `mime play`, `mime skit`, and `mime video`. For each search, we collected a maximum of 50 results, keeping only those videos that are between 1 and 10 minutes in duration and released under a Creative Commons license.

C.2 Model Settings and Parameters

We set the maximum output tokens for each model to be 128 tokens. We detail the settings of the models below.

- Gemini-1.5-Pro [25] and Gemini-2.5-Pro [12]: natively supports video as input, including audio.
- GPT-4o [1]: We sample 1 frame per second up to a maximum of 64 frames, in which case the frames are uniformly sampled. We resize the image to 512x512 to fit in the context window.
- Qwen2.5-VL-72B [61]: natively supports video as input, sampled at 2 frames per second for a maximum of 768 frames.
- LLaVA-Video-72B [89]: We sample 1 frame per second up to a maximum of 384 frames, in which case the frames are uniformly sampled.
- InternVL2.5 [10]: natively supports videos as input.
- VideoLLaMA3 [86]: natively supports video as input, with a maximum of 180 frames.

C.3 Supervised Finetuning

All fine-tuning experiments were conducted using the LLaMA Factory framework [93]. For the Qwen2.5-VL-72B model, we applied QLoRA fine-tuning with 8-bit quantization, a LoRA rank of 8, a learning rate of 1×10^{-4}, and trained for 3 epochs on MIMEQA. For the Qwen2.5-VL-7B model, we used standard LoRA fine-tuning with a rank of 8, a learning rate of 3×10^{-6}, and trained for 2 epochs. No additional hyperparameter tuning was performed.

C.4 Compute Resources

All evaluations and experiments in this paper were conducted on a remote cluster equipped with two NVIDIA H200 GPUs (each with 141 GB HBM3 memory). Runtime for each model evaluation varied between 1–10 hours depending on the model size.

C.5 Prompt Details

Below is the prompt template for Gemini-1.5-Pro, Gemini-2.5-Pro, and Qwen2.5-VL, which natively take in video input.

```
You are an expert in mime performance understanding and question answering.
Typically, the mime would use exaggerated gestures or pretend objects to convey a message.
Answer the question in one sentence using the video, with brief explanations.
Do not describe the frames just answer the question, and say nothing else.
If the mime is using imaginary objects, describe the objects as if they were real.
Question: <question>
```

As GPT-4o and LLaVA-Video require frames to be sampled from the video, we additionally specify the length of the video and the timestamps of the sampled frames in the text prompt. Below is the prompt template for GPT-4o and LLaVA-Video.

```
You are an expert in mime performance understanding and question answering.
Typically, the mime would use exaggerated gestures or pretend objects to convey a message.
The video lasts for {video_time}, and {num_frames} frames are uniformly sampled from it.
These frames are located at {frame_time}. Answer the question in one sentence using the video, with brief
explanations.
Do not describe the frames just answer the question.
If the mime is using imaginary objects, describe the objects as if they were real.
Question: <question>
```

Below is the prompt to GPT-4o for LLM-as-a-judge, which is adapted from Nagrani et al. [55].

```
Answer Grading Instructions:
Carefully consider the following question and answers regarding understanding of a mime performance.
You will be shown a "gold-standard" answer from a human annotator, referred to as the "Reference Answer", and a
"Candidate Answer".
Your task is to determine whether the candidate answer is a good answer in place of the "gold" reference using
the following criteria:

1. The candidate directly answers the question without deviation or misunderstanding.
2. The candidate does not contain misleading information and does not hallucinate story plots not present in
the reference answer.
3. Since the videos are mime performances, invisible actions, objects, or the mime actor portraying objects
should be considered correct if and only if they are relevant to the question.
4. The candidate answer can be a good answer in place of the reference answer even if they are not semantically
equivalent, as long as they are in the same ballpark, given that there can be multiple interpretations of a
mime acting out invisible actions or objects.

Your response should be one word, "TRUE" or "FALSE", and a brief explanation of your decision. You should
respond "TRUE" if the candidate is a good answer in place of the reference answer, and "FALSE" otherwise.
"""

GRADE_PROMPT = """
Question:
"{question}"
Candidate Answer:
"{candidate_answer}"
Reference Answer:
"{ref_answer}"
Equivalent? (True/False, and why?):
```

D Human Annotation Details

D.1 Guidelines for Annotators

Annotation Instructions

Each video ranges from 1 minute to 10 minutes. For each video, aim for approximately 6 scene-level questions, 4 global-level questions, and any relevant grounding questions. Mark explicit START and END timeframes for shot-level and local-level questions. Here is the question hiearchy:

Grounding This level mainly serves as a sanity check for whether the model understands the portrayed object in mime videos – what's the person doing, holding, and describing the imagined objects depicted.

- *E.g. What is the person in black shirt doing/holding/etc.?*

Scene-Level — local social information and temporal connection

- **Temporal** Interpreting sequences of events, causality, and the flow of actions within a specific timeframe.
  - **Template:** What caused (some event) to happen?
  - *E.g. What caused the person to fall over?*
  - *E.g. What happened before the person placed the spoon on the table?*

- **Affect Recognition** Tracking and analyzing emotions within a local scene, including subtle transitions and group sentiment.
  - **Template:** What is the attitude of (some person) towards (some event)?
  - **Template:** How does the (person) feel when (some event) happened?
  - *E.g. How is the person in black shirt feeling after placing the stone?*
  - *E.g. What is the attitude of the man towards the woman?*
  - *E.g. How did the group's emotional tone shift during the interaction?*
- **Intention and Behavior** Interpreting goals, plans, and motivations within a scene.
  - **Template:** Why did the (person) do (some action)?
  - *E.g. Why did the person holding the ice cream cry?*
  - *E.g. Why is the person in black outfit not speaking?*
  - *E.g. Why did the woman pretend not to notice the man?*
  - textitE.g. Why did the individual wait their turn before speaking?

Global-Level — focus on the overall narrative and high-level concepts

- **Working Memory** Retrieving, integrating, and reasoning with information across the entire video, beyond localized linear sequences. Requires the ability to decide relevance of information and present a coherent narrative.
  - **Template:** What happened after (some event)?
  - **Template:** How has the relationship between (person) and (person) changed?
  - **Template:** What would happen if (an event) didn't happen?
  - *E.g. What object in the beginning of the video foreshadowed the outcome?*
  - *E.g. How has the actions of the person changed throughout the video?*
  - *E.g. What event in the start of the video triggered the conflict in the final scene?*
- **Social Judgment** Evaluating behaviors, morality, and adherence to social norms, with consideration for cultural context and moral reasoning.
  - **Template:** How are the (person) and (person) getting along?
  - **Template:** How do the (person) actions demonstrate (social concept)?
  - *E.g. How does the person in the black shirt demonstrate rapport with the person in the blue dress?*
  - *E.g. What do the person's actions tell about his personality?*
  - *E.g. How might the group perceive this individual's behaviour?*
  - *E.g. How do the characters' behaviors suggest they are cooperating?*
- **Perspective Taking (Theory of Mind)** Inferring beliefs, desires, and emotions of others, including both cognitive and affective understanding.
  - **Template:** Does (person A) understand what (person B) was feeling?
  - **Template:** What is the (person) hoping to achieve?
  - **Template:** Would (person) do (action) after (event)?
  - *E.g. What goal does the main character pursue throughout the video?*
  - *E.g. How is the character's hope different from the reality?*
  - *E.g. Why is the main character motivated to change his behavior?*

D.2 Guidelines for Verifiers

Verification Instructions

This is a video question-answering dataset consisting of mime videos. The goal of this dataset is to evaluate whether current video language models can perform rich visual social reasoning without relying on natural language.

- Watch the entire video before reviewing the questions. The videos can be found [*link*]
- Answer each question based on the video content. If a question refers to a specific timestamp, focus on that section; otherwise, consider the whole video.
- Compare your answer with the "Reference Answer" column: Mark T in the "Answer Aligned" column if they align. ark F if they are clearly misaligned.
- For ambiguous questions, suggest a clearer version in the "Alternative Question" column.