SIV-Bench: A Video Benchmark for Social Interaction Understanding and Reasoning

Anonymous ACL submission

Abstract

Understanding social interaction, which encompasses perceiving numerous and subtle multimodal cues, inferring unobservable mental states and relations, and dynamically predicting others' behavior, is the foundation for achieving human-machine interaction. Despite rapid advances in Multimodal Large Language Models (MLLMs), the rich and multifaceted nature of social interaction has hindered the development of benchmarks that holistically evaluate and guide their social interaction abilities. Based on social relation theory, which has been widely regarded as a foundational framework for understanding social behavior, we provide SIV-Bench, a novel video benchmark for systematically evaluating MLLMs' capabilities across Social Scene Understanding (SSU), Social State Reasoning (SSR), and Social Dynamics Prediction (SDP). SIV-Bench features 2,792 originally collected video clips and 5,455 meticulously generated question-answer pairs derived from a human-LLM collaborative pipeline. It covers 14 typical relationships, diverse video lengths, genres, presentation styles, and linguistic and cultural backgrounds. Our comprehensive experiments show that leading MLLMs perform relatively well on SSU but remain weak on SSR and SDP, with the systematic confusion in relation inference as a key bottleneck. An in-depth analysis of the reasoning process attributes MLLMs' suboptimal performance to misalignment with human thoughts and insufficient reasoning depth. Moreover, we find audio and subtitles aid in reasoning-intensive SSR and SDP. Together, SIV-Bench offers a unified testbed to measure progress, expose limitations, and guide future research toward more socially intelligent MLLMs.

1 Introduction

The rapid development of Multimodal Large Language Models (MLLMs) capable of processing text, images, and video has driven strong performance across tasks such as visual reasoning, video captioning, and multimodal dialogue (Team et al., 2023; Wu et al., 2024; Zhu et al., 2025b; Hurst et al., 2024; Yang et al., 2024). As these capabilities expand, there is a growing need for benchmarks that can evaluate model performance, uncover limitations, and guide future research (Fu et al., 2024; Fang et al., 2024; Zhou et al., 2024; Qiang et al., 2025; Huang et al., 2024). One critical yet underexplored area is the social interaction understanding and reasoning, which is a core aspect of social intelligence that encompasses not only observable behaviors but also implicit mental states and social relationships governing behaviors such as forming bonds, exchanging information, and coordinating actions (Berger et al., 1972; Smith-Lovin and Heise, 1988). However, existing video benchmarks, whether designed for specific tasks like video object segmentation (Ding et al., 2025; Athar et al., 2025), captioning (Chen et al., 2025; Wu et al., 2025), and fine-grained action understanding (Perrett et al., 2025), or for broader video understanding (Wang et al., 2025; Li et al., 2024b,c), still struggle to systematically probe MLLMs' understanding of the multifaceted nature of social interaction.

To address this gap, we *first* decompose the capacity to understand and reason about social interaction into three core, interrelated dimensions. **1) Social Scene Understanding (SSU)** is foundational, enabling the recognition of visible elements such as objects, environments, and socially salient human features like body movements, clothing, and physical appearance. Reliable scene perception is required to ground interpretations in relevant cues. **2) Social State Reasoning (SSR)** is essential for interpreting the unobservable states of interaction, such as emotions, intents, attitudes, and interpersonal relationships, which guide and shape behavior (Strachan et al., 2024; Wu et al., 2020). This capacity allows models to move beyond surface-level

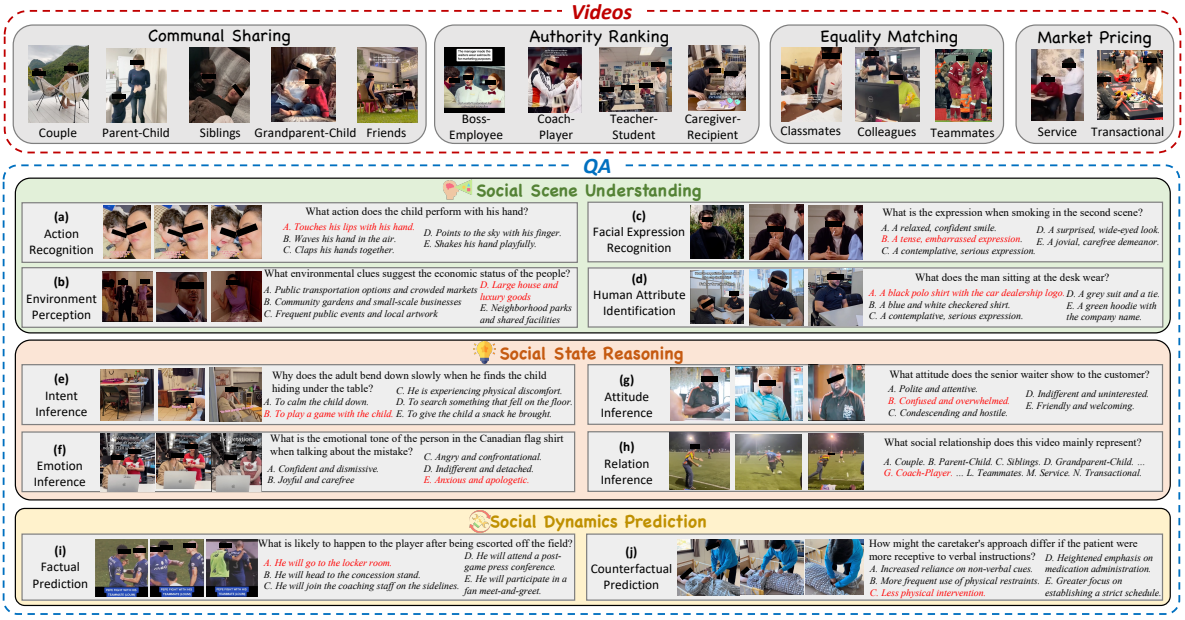


Figure 1: Overview of SIV-Bench, showing its diverse videos spanning various social interactions and sample QAs for three task dimensions: Social Scene Understanding (SSU), Social State Reasoning (SSR), and Social Dynamics Prediction (SDP), along with their fine-grained sub-tasks.

features and grasp the underlying states. **3) Social Dynamics Prediction (SDP)** enables the model to reason about how interactions proceed over time or under alternative conditions, capabilities essential for a flexible and human-like understanding of social scenarios (Ramnani and Miall, 2004; Byrne, 2016). It involves factual prediction and counterfactual prediction. The former anticipates upcoming actions or changes in social states, while the latter examines how alterations in social scenes or states can affect interaction outcome.

Second, to systematically and comprehensively evaluate the three dimensions, we introduce **SIV-Bench** (**S**ocial **I**nteraction **V**ideo **Bench**mark) grounded in social relation theory, which has been widely regarded as a foundational framework for understanding social behavior (Thibaut, 2017; Burkitt, 1997; Hartup, 1989). Specifically, SIV-Bench is built on Fiske's Relational Models Theory (Fiske, 1992), categorizing social interactions via four foundational models (*Communal Sharing*, *Authority Ranking*, *Equality Matching*, and *Market Pricing*), instantiated through 14 relation types (e.g., parent-child, friends, colleagues). This relational context underpins all three dimensions. It conditions how cues are interpreted in SSU (e.g., a gaze between colleagues vs. lovers), modulates mental-state inference in SSR (e.g., criticism from a mentor vs. a stranger), and shapes SDP by constraining future behaviors and counterfactuals through relational norms and history (e.g., siblings vs. business rivals). SIV-Bench comprises 2,792 video clips sourced from TikTok and YouTube, and a QA pipeline that combines adversarial filtering with human verification produces 5,455 high-quality questions. To systematically assess the contribution of linguistic cues, SIV-Bench further provides audio tracks and three subtitle conditions: the original version (origin), a version with transcribed dialog added (+sub), and a version with all on-screen text removed (-sub).

In the experiments, we evaluate a broad set of models, including leading commercial MLLMs, strong open-source MLLMs, and video-specialized models. Overall, models are comparatively strong on SSU but weak on high-level reasoning, particularly within SSR, where **relation inference** exhibits systematic confusions. Our fine-grained analyses reveal recurring error sources, including selecting plausible but secondary relations, over-reliance on contextual cues, insufficient commonsense social reasoning, and missed perceptual details. SDP remains challenging as well, with counterfactual cases being relatively better handled by top models. We further study textual cues via controlled subtitle ablations and observe a task-dependent effect. Removing language typically has a limited impact on basic perception, but can degrade performance on inference-intensive SSR/SDP cases. Furthermore, on a hard diagnostic subset with brief explanations

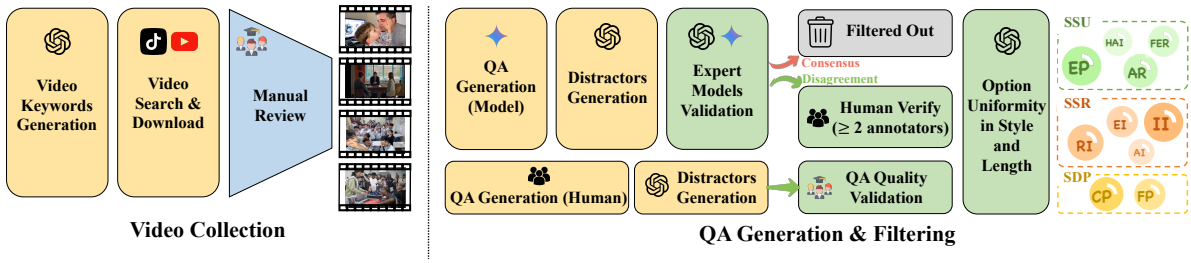


Figure 2: The SIV-Bench construction pipeline, detailing the data collection process (left), and the QA generation & filtering process (right) with human-LLM collaboration. In the diagram, blocks indicate content (like keywords, video and QA) generation steps, blocks represent validation or modification stages.

and a human baseline, we observe a large remaining gap to humans, underscoring the difficulty of robust, human-aligned social reasoning.

Our key contributions are as follows: **1)** We propose a novel analytical framework that structurally decomposes the complex task of multimodal social interaction understanding and reasoning into three core, interrelated dimensions, each further detailed into fine-grained sub-tasks. **2)** We introduce SIV-Bench, a new video benchmark specifically curated for the analysis and comprehension of complex real-world social interactions. SIV-Bench comprises 2,792 real-world video clips representing 14 distinct social relationship types, and features 5,455 high-quality question-answer pairs generated through a human-LLM collaborative pipeline. **3)** Our comprehensive experiments on diverse MLLMs reveal the limitations in their current capacity for deep human social understanding.

2 SIV-Bench

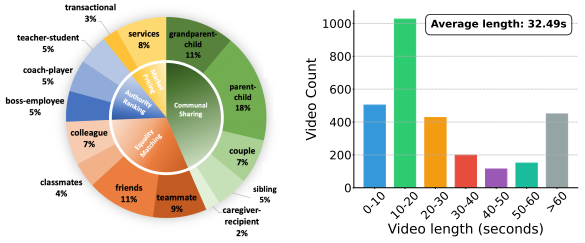
This section presents the construction of SIV-Bench, including video collection (Section 2.1), QA composing (Section 2.2), and a comparison with existing benchmarks (Section 2.3). Figure 1 offers some illustrative examples from SIV-Bench. Figure 2 shows the construction pipeline.

2.1 Video Collection

Firstly, we utilize GPT-4o-mini (Hurst et al., 2024) to generate comprehensive search keywords for each of the 14 relationship types (the word-clouds are shown in Figure 8), specifically including terms associated with varying degrees of intimacy, both positive (e.g., "love", "encourage") and negative (e.g., "conflict", "fight"). Leveraging these keywords, we conduct targeted searches and download initial video candidates from TikTok and YouTube platforms using Python libraries such as `TikTokApi` and `yt-dlp`, yielding approximately 5000 raw video clips. Each video is then manually reviewed by the authors to ensure it contains clearly observable and meaningful social interactions. Videos are excluded if they do not depict clear social interaction (e.g., a vlogger speaking directly to the camera without interacting with others), if the interaction context does not fit within a set of well-defined interpersonal scenarios (e.g., an interview setting with scripted dialogue), or if the dominant interaction is difficult to identify due to the presence of multiple overlapping social dynamics (e.g., a large multi-generational family posing for a group photo). These criteria are designed to ensure that each included video primarily features one interpretable and coherent type of interpersonal interaction, allowing for more consistent analysis.

The SIV-Bench comprises **2,792** curated video clips, with statistics shown in Figure 3. The collection showcases a rich diversity in social relationships (Figure 3a). Communal Sharing interactions are the most represented category, reflecting the naturalistic prevalence and psychological centrality of these relationships in daily life (Simão and Seibt, 2014; Kameda et al., 2005). The other three relational models are also well represented, ensuring broad interpersonal coverage. In addition to relationship diversity, the benchmark covers heterogeneous genres and filming/editing conventions, which helps test models under different visual narratives and interaction realizations. In terms of duration (Figure 3b), clips average 32.49 seconds. While most clips are 10–20 seconds, the dataset spans a wide distribution, including many short clips (under 10 seconds) and a significant number over 60 seconds. English predominates, but SIV-Bench also includes other languages, adding to multicultural diversity (detailed in Figure 9). Details on video diversity are in Appendix A.2.

To evaluate how different forms of textual information affect MLLMs' understanding of social



(a) Relation type distribution. (b) Video length distribution.

Figure 3: Video statistics for SIV-Bench.

interactions, we implement specific subtitle processing methods (Figure 4). Many original videos ('Origin') contain embedded on-screen text that often serves as scene descriptors or keywords (e.g., the 'Buy and Sell' text overlay). To focus on visual and auditory cues, we create a '-Subtitle' version by removing such original textual overlays using `video-subtitle-remover` (YaoFANGUK, 2025). Conversely, to provide full access to spoken dialogue, we generate a '+Subtitle' version. We employ Whisper-large-v3 (Radford et al., 2022) for audio transcription of the dialogue, and then use GPT to translate these transcriptions into English, ensuring consistent and high-quality subtitles (e.g., 'He sold us his iPhone 12 with 256GB').

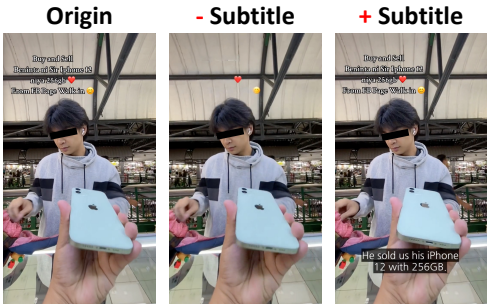


Figure 4: Illustration of the three subtitle conditions.

2.2 QA Composing

Our Question-Answer (QA) composition pipeline is designed to move beyond simple recognition tasks and ensure a high density of challenging reasoning problems. The process consists of three main stages: scalable generation, adversarial filtering, and human-in-the-loop curation.

Scalable Generation. We begin by generating a large initial pool of diverse QA pairs for each video, leveraging the capabilities of Gemini-2.0-Flash with full video input. The prompts are carefully designed to elicit questions that span our SSU, SSR, and SDP evaluation dimensions. This stage typically yields an average of 10 QA pairs per video. Subsequently, we employ GPT-4o-mini to generate four distractors for each QA pair, ensuring they are contextually relevant yet clearly distinguishable from the correct answer. Separating distractor generation significantly reduces ambiguity compared to simultaneous generation.

Adversarial Filtering. To ensure SIV-Bench serves as a rigorous diagnostic tool rather than a saturated test set, we implement a strict *Model-Consensus Filtering* mechanism. We utilize three commercial models (Gemini-2.0-Flash, Gemini-2.0-Pro, and GPT-4o-mini) to independently answer all candidate questions. We adopt an adversarial approach: questions correctly answered by all models (indicating trivial reasoning or visual obviousness) are discarded. We focus exclusively on the remaining "hard" samples where model consensus is not achieved, as these disagreements signal the need for deeper social reasoning.

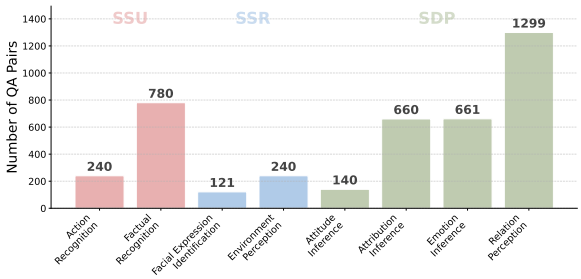


Figure 5: QA Distribution across 10 sub-tasks.

Human Verification and Authoring. The questions surviving the adversarial filter undergo a rigorous human-in-the-loop process to distinguish legitimate challenging samples from noise: **(1)** We recruit 20 human annotators to verify the non-consensus items. Each QA is reviewed by at least two annotators. Only questions where humans independently agree on the correct answer are retained. This yields 3,096 high-quality, model-challenging QAs. **(2)** To further expand the benchmark's ceiling, annotators are instructed to generate novel, complex questions and their answers. These human-authored 2,359 QAs are designed to probe nuances often missed by model generators. Appendix B.2 details the annotation guidelines and interface. For all new human-generated QAs, distractors are created using GPT-4o-mini.

Finally, all curated QA pairs undergo an automated refinement stage to standardize linguistic style and option length, minimizing superficial cues that models might exploit. Full prompts are provided in Appendix B.1. Through this "Filter-

Table 1: Comparison of various benchmarks, including total number of items (**#Items**, representing the number of videos, dialogues, images, etc.), number of QA pairs (**#QAs**), annotation method (**Anno.**, M/A means manually/automatic manner), and tasks (*Task Types*: **SU** for Scene Understanding, **SR** for State Reasoning, **DP** for Dynamics Prediction). Note that *Task Types* here refer to general task categories, which include both social and physical scenarios. The table also shows whether each benchmark includes **Multi-Person Interaction**, covers **Various Relations**, is based on newly collected data (**Original Collection**), and provides subtitle/audio (**S.A.**).

Benchmark	#Items	#QAs	Anno.	*Task Types*			Multi-Person	Various	Original	S.A.
				SU	SR	DP	Interaction	Relations	Collection	
Social Relation Inference										
DialogRE (Yu et al., 2020)	1,788	-	M	✓	✓	✗	✓	✓	✗	-
PIPA (Sun et al., 2017)	37,107	-	M	✗	✓	✗	✓	✓	✗	-
ViSR (Liu et al., 2019)	8,000	-	M	✗	✓	✗	✓	✓	✓	✗
General Video Understanding and Reasoning										
VideoMME (Fu et al., 2024)	900	2,700	M	✓	✓	✗	✓	✗	✓	✓
MLVU (Zhou et al., 2024)	1,730	3,102	M	✓	✗	✓	✓	✗	✗	✗
VideoVista (Li et al., 2024c)	894	24,906	A	✓	✗	✓	✗	✗	✗	✓
Social-IQ 2.0 (Wilf et al., 2023)	1,000	6,000	M	✓	✓	✗	✓	✓	✓	✗
Social Genome (Mathur et al., 2025)	272	1,486	M	✓	✓	✗	✓	✗	✗	✓
Perception Test (Patraucean et al., 2023)	11,620	38,000	M	✓	✗	✓	✗	✗	✓	✗
MVBench (Li et al., 2024b)	3,641	4,000	A	✓	✗	✓	✗	✗	✗	✗
Video-Bench (Ning et al., 2023)	5,917	17,036	A&M	✓	✓	✗	✓	✗	✗	✗
EgoSchema (Mangalam et al., 2023)	5,063	5,063	A&M	✓	✓	✗	✓	✗	✗	✗
SIV-Bench	2,792	5,455	A&M	✓	✓	✓	✓	✓	✓	✓

then-Verify" pipeline, SIV-Bench ultimately comprises **5,455** high-quality QA pairs. The distribution across core dimensions is illustrated in Figure 5, featuring a strong emphasis on SSR and SDP, enabling rich evaluation of higher-order social intelligence. Appendix B.3 provides detailed statistics confirming structural balance and diversity.

2.3 Comparison with Existing Benchmarks

Table 1 highlights the key differences between SIV-Bench and existing works. Prior datasets on social interaction, such as DialogRE (Yu et al., 2020) (text-only), PIPA (Sun et al., 2017) (image-only), and ViSR (Liu et al., 2019), are limited to relation recognition, single modalities, and lack task diversity and scalability due to manual annotation. Currently, most general video understanding benchmarks (like VideoVista (Li et al., 2024c) and MVBench (Li et al., 2024b)) lack a focus on social interaction; others (like MLVU (Zhou et al., 2024) and Video-Bench (Ning et al., 2023)), while touching upon it, do not fully cover social relations. Even Social-IQ 2.0 (Wilf et al., 2023), which concentrates on this area, has limitations in task diversity and dynamic reasoning. SIV-Bench is built on original data, combines manual and automatic annotations, and is one of the few benchmarks to provide subtitle and audio information, supporting richer multimodal social reasoning.

3 Experiments

3.1 Experimental Setup

We evaluate a diverse set of closed- and open-source MLLMs on SIV-Bench, including Gemini-2.0/2.5-Flash (Google, 2025a,b), Gemini-2.5-Pro (Doshi, 2025), GPT-4o (Hurst et al., 2024), o4-mini (OpenAI, 2025), Qwen2.5-VL-7B/72B-Instruct (Bai et al., 2025), mPLUG-Owl3 (Ye et al., 2024), InternVL3-8B/78B (Zhu et al., 2025a), LLaVA-OneVision (Li et al., 2024a), and LLaVA-Video (Zhang et al., 2024). Evaluations are conducted using VLMEvalKit (Duan et al., 2024), which provides a unified evaluation interface across MLLMs. All models generate responses under their default inference settings to ensure a fair comparison.

We use a standardized prompt (Figure 18) across all models, asking for both an option letter (e.g., 'A.') and the full answer text. To parse outputs robustly, we first extract a valid option letter; if missing, we match the raw output to all options using text similarity. To enable scalable and model-agnostic evaluation across heterogeneous MLLMs, we cast all tasks into a unified multiple-choice interface and report accuracy as the primary metric.

3.2 Overall Performance

The results shown in Table 2 highlights a clear performance stratification primarily driven by model scale. While the proprietary Gemini-2.5-Pro

Table 2: Evaluation results of MLLMs on SIV-Bench. Accuracy (%) on SSU, SSR, SDP, and Overall under different subtitle settings ('origin', '+sub', '-sub'). Statistical significance tests are reported in Appendix C.4.

Models	Params	Social Scene Understanding			Social State Reasoning			Social Dynamics Prediction			Overall		
		origin	+ sub	- sub	origin	+ sub	- sub	origin	+ sub	- sub	origin	+ sub	- sub
Open-source MLLMs													
mPLUG-Owl3	7B	46.11	45.94	46.34	39.78	39.50	38.13	44.30	46.08	44.35	42.06	42.42	41.15
LLaVA-OneVision	7B	39.20	39.41	40.08	41.95	43.65	38.66	43.79	44.51	39.42	41.97	43.04	39.42
LLaVA-Video	7B	50.22	50.61	50.56	39.33	38.19	36.14	41.60	42.14	39.94	41.09	42.00	38.66
Qwen2.5-VL-7B-Instruct	7B	51.22	50.88	50.22	40.24	38.94	37.66	42.69	43.82	42.24	44.02	44.21	41.65
InternVL3-8B	8B	56.83	56.13	56.50	40.35	40.90	37.92	44.53	45.52	44.40	45.82	46.05	44.56
Qwen2.5-VL-72B-Instruct	72B	75.73	76.24	73.54	52.25	52.75	51.21	59.02	58.40	57.78	58.80	59.63	57.66
InternVL3-78B	78B	71.46	73.66	71.76	51.65	52.39	50.14	55.77	56.28	54.25	55.46	56.32	54.50
Closed-source MLLMs													
o4-mini	-	78.83	79.04	78.13	50.47	51.30	48.99	56.89	56.00	55.26	55.68	55.89	54.54
GPT-4o	-	79.10	79.74	78.06	52.73	53.20	51.79	59.02	60.59	58.60	58.02	58.86	56.99
Gemini-2.0-Flash	-	78.46	78.16	78.34	51.89	52.43	49.78	57.59	58.63	55.70	56.40	57.23	54.64
Gemini-2.5-Flash	-	81.70	82.14	79.71	48.99	50.54	47.60	59.47	59.95	56.88	57.87	58.11	56.05
Gemini-2.5-Pro	-	**85.07**	**85.41**	**84.94**	**54.30**	**54.85**	**52.32**	**60.45**	**61.54**	**58.83**	**61.65**	**62.40**	**60.22**

achieves the state-of-the-art (62.40% w/ subtitles), large-scale open-source models demonstrate exceptional competitiveness; notably, Qwen2.5-VL-72B (59.63%) outperforms other leading proprietary systems like GPT-4o and Gemini-2.5-Flash (>58%). A significant capacity gap is evident, as larger models consistently dominate their smaller counterparts (e.g., Qwen-72B's 59.63% vs. 7B's 44.21%), suggesting that sufficient parameter count is a prerequisite for robust social reasoning. We verify that the observed performance gaps are statistically significant using paired bootstrap tests; full details are reported in Appendix C.4.

3.3 Decomposed Performance

Table 3: Relative performance change (%) against the baseline. The audio ablation (w/o Audio) is evaluated on Gemini-2.5-Flash, while subtitle variations (+Sub/-Sub) reflect the average impact across all models.

Condition	SSU	SSR	SDP	Overall
w/o Audio	-0.35	-2.44	-2.33	-1.40
+ Subtitle	+0.05	+0.28	+0.65	+0.15
- Subtitle	-0.97	-2.07	-1.68	-0.90

Subtitle and Audio Influence. To assess the impact of different modalities, we analyze the relative performance changes compared to the original video baseline, as shown in Table 3. The audio ablation is conducted on Gemini-2.5-Flash, while subtitle variations represent the average results across all evaluated models. The results demonstrate a clear hierarchy of modality importance. Removing audio leads to the most substantial performance drop (-1.40%), particularly in reasoning-intensive tasks (SSR: -2.44%), highlighting the necessity of auditory cues for social interaction understanding and reasoning. For linguistic cues, the impact is notably asymmetric: while adding subtitles provides only marginal gains (+0.15%), their removal significantly hinders performance (-0.90%), especially in SSR and SDP. This suggests that while models may not always benefit from redundant text, the absence of critical textual context creates a significant bottleneck for complex social reasoning.

Task Influence. SSU focuses on recognizing visible elements and is generally the easiest dimension, where models achieve their highest scores. This is not always the case, since LLaVA-OneVision sometimes performs better on SSR or SDP. In SSU, larger models, including closed-source models and high-parameter open-source models, consistently outperform smaller ones, which likely reflects stronger perceptual ability from greater capacity and broader training (Alabdulmohsin et al., 2022). After SSU, SDP is usually the next most tractable, while SSR remains the most difficult because it requires inferring latent social states such as emotions and intentions. Figure 6 further breaks down performance across 10 fine-grained tasks in original videos. Stronger closed-source models such as the Gemini and GPT-4o form the outer ring and outperform smaller open-source models. In SSU, the performance gap is most pronounced in Action Recognition (AR) and Facial Expression Recognition (FER), suggesting advantages in capturing subtle visual cues. SSR is more challenging. Models perform moderately on Intent Inference (II) and Emotion Inference (EI), but they struggle most with Relation Inference (RI), which is often the lowest point on the radar. In

SDP, performance improves on Factual Prediction (FP) and Counterfactual Prediction (CP), which may reflect social commonsense acquired from language data. Most models perform better on CP than FP, possibly because hypothetical framing provides clearer reasoning cues. Representative failures are shown in Figure 19, 20, and 21.

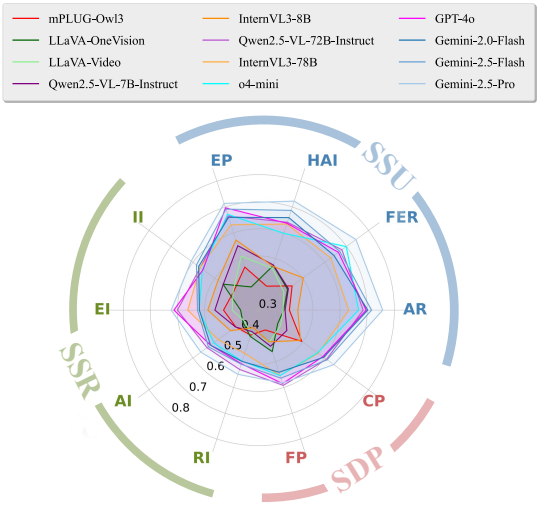


Figure 6: Radar chart of MLLM performance across the 10 fine-grained SIV-Bench sub-tasks.

3.4 In-Depth Analysis of Relation Inference

Since RI emerges as a key bottleneck and provides the relational grounding needed to interpret many higher-level social states, we take a closer look at RI and characterize its error structure quantitatively. Figure 7 aggregates RI predictions across the evaluated models into a confusion matrix, which reveals structured error clusters rather than random noise: a prominent pattern is Authority–Equality confusion, where hierarchical relations such as Boss–Employee and Coach–Player are frequently predicted as their egalitarian counterparts (Colleagues and Teammates). Errors also concentrate within relation families. The red dashed boundaries delineate the four foundational relational models, and off-diagonal clusters around these boundaries highlight systematic confusions between conceptually adjacent relation groups.

Qualitatively, we observe four common failure modes. **(1)** Models fail to differentiate primary vs. secondary relations in multi-relational scenarios, predicting a plausible but less salient relation. **(2)** Models are misled by scene- and human-induced cues, over-relying on stereotypical settings or surface language without validating the actual interaction (e.g., classroom → teacher–student). **(3)** Some errors reflect deficient commonsense social reasoning, where predictions contradict basic social conventions (e.g., two <friends> wearing different sports team uniforms but showing friendly behavior are misclassified as <teammates>, disregarding the fact that such uniforms typically signify competition). **(4)** Models ignore key perceptual details (e.g., age/role cues) that are decisive for disambiguating closely related relations. More failure cases are provided in Appendix C.2.

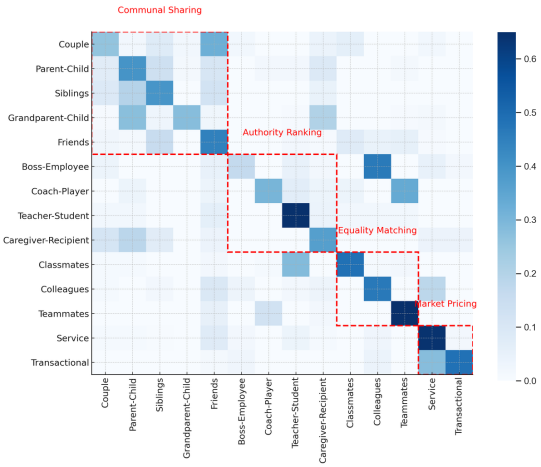


Figure 7: Aggregated Confusion Matrix for the Relation Inference (RI) task across all models. Red dashed lines delineate the four foundational relational models.

3.5 In-depth Analysis of Reasoning Process

To diagnose the cognitive boundaries of current MLLMs, we establish SIV-Bench-Hard, a subset of 200 QAs where models collectively demonstrate the highest error rates. This focused selection enables concentrated failure-mode analysis and facilitates a reliable human baseline, which is prohibitively expensive at full scale. We recruit three independent crowdworkers, distinct from the original annotators, to provide answer selections and free-text explanations. Table 4 highlights a substantial performance gap, with Gemini-3-Pro and GPT-5.1 achieving accuracies of 45.50% and 39.00%, compared to a human baseline of 74.40%.

LLM-judge results show a gap between reasoning presentation and social cognition. Models achieve strong structural scores (Logical Coherence, Relevance >4.0) but substantially lower Alignment and Depth (<3.5). We also find that reasoning similarity to human traces is positively associated with answer correctness (Figure 24, Table 9), proving that robust social interaction understanding

Table 4: Accuracy and LLM-judge scores (1-5) on reasoning quality (Rel=Relevance, Cohe=Logical Coherence, Depth=Depth of Analysis, Align=Alignment with Human, Conc=Conciseness, Ovrl=Overall).

Model	Acc%	Rel	Align	Cohe	Depth	Conc	Ovrl
Human	74.40	-	-	-	-	-	-
Gemini-3-Pro	**45.50**	**4.66**	**3.30**	**4.67**	**3.49**	4.87	**4.10**
GPT-5.1	39.00	4.58	3.29	4.65	3.26	4.88	4.00
Gemini-2.5-Pro	37.00	4.57	3.26	4.65	3.41	4.89	4.05
Gemini-2.5-Flash	32.32	4.48	3.17	4.55	3.22	4.87	3.95
GPT-4o-mini	29.00	4.45	3.20	4.56	3.12	**4.91**	3.90
Qwen2.5-VL-7B	24.50	4.00	2.89	4.21	3.05	4.45	3.63

requires human-like inference rather than purely fluent justifications. Interestingly, while Gemini-3-Pro maintains a lead in analysis depth over GPT-5.1 (+0.23), this does not improve alignment, indicating that an increased depth of reasoning does not inherently guarantee social accuracy.

A primary failure mode involves the inability to identify core social elements. In 40% to 54% of failures, models prioritize secondary visual cues over the "long-term separation" or "familial bonds" emphasized by humans. Furthermore, 45% of depth-related failures stem from a lack of social dynamic analysis, where models stop at direct causality and ignore power structures or cultural norms. We also observe a unique "excessive caution" bias in GPT-5.1, which frequently refuses to make valid socio-cultural inferences from appearance, resulting in the lowest alignment scores where humans readily use such cues for grounding. These insights suggest that MLLMs require multi-layered social grounding that explicitly partitions inference into observational, affective, and normative layers, enabling the models to bridge the gap between literal perception and human-aligned social intelligence.

4 Related Works

Video Benchmarks for MLLMs. Rapid progress in video MLLMs has motivated a growing suite of benchmarks spanning perception to high-level reasoning. Representative benchmarks cover multiple tasks (e.g., QA and summarization) over diverse video sources (Zhou et al., 2024; Fu et al., 2024; Li et al., 2024b; Fang et al., 2024; Ning et al., 2023). A major line of work emphasizes long-form understanding and temporal reasoning (Rawal et al., 2024; Mangalam et al., 2023; Wang et al., 2024b; Liu et al., 2024). Meanwhile, emerging benchmarks probe new deployment regimes, such as real-time/streaming video reasoning ((Xun et al., 2025)) and egocentric intent understanding with gaze cues ((Peng et al., 2025)). While these benchmarks provide broad coverage of video comprehension, they are not designed to systematically evaluate the fine-grained understanding and reasoning required for complex multi-person social interactions.

Evaluating Social Intelligence in AI Systems. Beyond general video understanding, a parallel line of work evaluates social intelligence, moving from component-level signals (e.g., actions and affect) toward richer interpersonal reasoning and grounded social cognition. Recent benchmarks provide narrower but deeper probes, including Theory-of-Mind understanding from short films (Villa-Cueva et al., 2025), grounded social reasoning traces from interaction videos (Mathur et al., 2025), and non-verbal social reasoning without speech (Li et al., 2025). Complementary efforts assess emotional and social intelligence in video QA (Zhang et al., 2025) and social intelligence in authentic multi-turn human conversations (Huang et al., 2025), alongside interactive social environments (Zhou et al., 2023; Wang et al., 2024a). However, existing benchmarks largely evaluate social understanding in a task-specific or modality-specific manner. SIV-Bench complements them by providing a unified evaluation of social interaction understanding in real-world videos, spanning perception, social state reasoning, and dynamics prediction.

5 Conclusion

We introduce **SIV-Bench**, a video benchmark for evaluating MLLMs on real-world social interaction understanding and reasoning grounded in foundational social relationship models. SIV-Bench contains 2,792 videos and 5,455 curated QA pairs, organized into three dimensions (SSU, SSR, and SDP) as well as ten fine-grained sub-tasks for diagnosis. Extensive experiments show that current MLLMs perform relatively well on SSU but struggle with high-level reasoning, with Relation Inference exhibiting systematic confusions across relation types. Subtitle ablations suggest that textual signals can affect performance in a task-dependent manner, particularly for complex inference cases. On a challenging hard subset, we further observe a substantial gap to human performance, reinforcing the difficulty of human-aligned social reasoning. We hope SIV-Bench will serve as a unified testbed to track progress and drive more robust social intelligence in MLLMs.

Limitations

While SIV-Bench offers a structured diagnostic benchmark for social interaction understanding, the current size and coverage still leave room to expand toward broader social settings, cultures, and interaction styles. Moreover, our primary evaluation remains multiple-choice, which simplifies scoring but does not fully capture open-ended grounding, generation, or interactive behaviors. We include SIV-Bench-Hard with brief explanations to partially address this gap, yet it is relatively small and still constrained in evaluation scope. Finally, social interpretation can be subjective and culturally contingent, and some clips may admit multiple plausible readings, suggesting the need for future work on more calibrated human protocols and richer evaluation formats.

Ethical Considerations

SIV-Bench is built from publicly available videos (e.g., TikTok and YouTube), and we carefully consider licensing and privacy risks when creating and releasing the benchmark. To reduce copyright and platform-policy concerns, we will not redistribute raw video files; instead, we will release annotations along with video identifiers (e.g., URLs and timestamps) and provide a script for users to retrieve content directly from the original platforms, following common practice in prior video benchmarks.

Because the videos may contain identifiable individuals, we applied manual filtering to exclude content that appears overly private or sensitive, and we encourage responsible downstream use. Our human annotation procedures were designed to be minimal-risk, with annotators compensated and handled anonymously. Finally, while SIV-Bench is intended to support transparent evaluation of MLLMs' social understanding, we acknowledge the dual-use nature of social inference technologies and recommend that users avoid deployment in high-stakes settings and follow applicable privacy and platform guidelines when using the benchmark.

References

Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. 2022. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312.

Ali Athar, Xueqing Deng, and Liang-Chieh Chen. 2025. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19023–19035.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Joseph Berger, Bernard P Cohen, and Morris Zelditch Jr. 1972. Status characteristics and social interaction. *American sociological review*, pages 241–255.

Ian Burkitt. 1997. Social relationships and emotions. *Sociology*, 31(1):37–55.

Ruth MJ Byrne. 2016. Counterfactual thought. *Annual review of psychology*, 67(1):135–157.

Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. 2025. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. *arXiv preprint arXiv:2502.12782*.

Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. 2025. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*.

Tulsee Doshi. 2025. Build rich, interactive web apps with an updated gemini 2.5 pro. https://blog.google/products/gemini/gemini-2-5-pro-updates/. Accessed: 2025-05-09.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Alan P Fiske. 1992. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Google. 2025a. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/. Accessed: 2025-05-09.

Google. 2025b. Start building with Gemini 2.5 Flash. Google for Developers Blog. Accessed: May 9, 2025.

Willard W Hartup. 1989. Social relationships and their developmental significance. *American psychologist*, 44(2):120.

Shuai Huang, Wenxuan Zhao, and Jun Gao. 2025. Si-bench: Benchmarking social intelligence of large language models in human-to-human conversations. *arXiv preprint arXiv:2510.23182*.

Yizhe Huang, Xingbo Wang, Hao Liu, Fanqi Kong, Aoyang Qin, Min Tang, Xiaoxi Wang, Song-Chun Zhu, Mingjie Bi, Siyuan Qi, and 1 others. 2024. Adasociety: An adaptive environment with social structures for multi-agent decision-making. *arXiv preprint arXiv:2411.03865*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tatsuya Kameda, Masanori Takezawa, and Reid Hastie. 2005. Where do social norms come from? the example of communal sharing. *Current Directions in Psychological Science*, 14(6):331–334.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Hengzhi Li, Megan Tjandrasuwita, Yi R Fung, Armando Solar-Lezama, and Paul Pu Liang. 2025. Mimeqa: Towards socially-intelligent nonverbal foundation models. *arXiv preprint arXiv:2502.16671*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024c. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3566–3574.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

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

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

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. Accessed: 2025-05-09. The direct PDF can also be found at https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. 2025. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, and 1 others. 2025. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913.

Rushi Qiang, Yuchen Zhuang, Yinghao Li, Dingu Sagar V K, Rongzhi Zhang, Changhao Li, Ian Shu-Hei Wong, Sherry Yang, Percy Liang, Chao Zhang, and Bo Dai. 2025. Mle-dojo: Interactive environments for empowering llm agents in machine learning engineering. *Preprint*, arXiv:2505.07782.

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

Narender Ramnani and R Christopher Miall. 2004. A system in the human brain for predicting the actions of others. *Nature neuroscience*, 7(1):85–90.

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

Cláudia Simão and Beate Seibt. 2014. Gratitude depends on the relational model of communal sharing. *PloS one*, 9(1):e86158.

Lynn Smith-Lovin and David R Heise. 1988. Analyzing social interaction. *Advances in affect control theory. Gordon and Breach*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A domain based approach to social relation recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3481–3490.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

John W Thibaut. 2017. *The social psychology of groups*. Routledge.

Emilio Villa-Cueva, SM Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. 2025. Moments: A comprehensive multimodal benchmark for theory of mind. *arXiv preprint arXiv:2507.04415*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024a. Sotopia-π: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and 1 others. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

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

Haiyan Wu, Xun Liu, Cindy C Hagan, and Dean Mobbs. 2020. Mentalizing during social interaction: A four component model. *Cortex*, 126:242–252.

Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. 2025. Event-equalized dense video captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8417–8427.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, and 1 others. 2025. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *arXiv preprint arXiv:2505.02064*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

YaoFANGUK. 2025. video-subtitle-remover. https://github.com/YaoFANGUK/video-subtitle-remover. Accessed: 2025-05-09.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. *arXiv preprint arXiv:2004.08056*.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, and 1 others. 2025. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Video Details

A.1 Video Collection

To compile a rich and diverse video corpus for SIV-Bench that reflects a wide array of social interactions, we first identify 14 key social relationship types (e.g., parent-child, friends, colleagues, service interactions). For each of these types, we employ GPT-4o mini to generate a comprehensive set of search keywords. This strategy is designed to ensure broad coverage in our video collection from platforms like YouTube and TikTok, capturing various interaction categories (such as cooperation, conflict, and support), forms (including verbal discussions, non-verbal expressions, and shared activities), and differing levels of intimacy or formality inherent to each relationship. Figure 8 visualizes these tailored keyword sets as word clouds for each of the 14 relationship types, illustrating the breadth of concepts targeted. This targeted yet diverse keyword generation is crucial for assembling a video dataset that truly represents the complexity of human social life.

A.2 Video Diversity

The SIV-Bench video corpus is intentionally diverse in both subject matter and presentation to ensure comprehensive model evaluation. This genre diversity, illustrated in Figure 10, includes

(a) Grandparent–Child (b) Parent–Child

(c) Siblings (d) Couple

(e) Friends (f) Boss–Employee

(g) Coach–Player (h) Teacher–Student

(i) Caregiver–Recipient (j) Classmates

(k) Teammates (l) Colleague

(m) Transactional (n) Service

Figure 8: Word clouds of GPT-generated keywords used for sourcing videos across 14 distinct social relationship types. Keywords capture diverse interaction scenarios, roles, and intimacy levels from platforms such as YouTube and TikTok.

a wide array of video types such as candid Daily Life recordings, scripted Movie Clips, dynamic Sports Replays, illustrative Commercials, and non-photorealistic Animated Videos. Complementing this, Figure 11 showcases varied visual presentation styles, ranging from traditional Third-Person perspectives and immersive First-Person (POV) footage to interactions depicted via Phone Call interfaces and creative Solo Multi-Role Sketches. This multifaceted diversity in video content and form provides a robust and varied testbed.

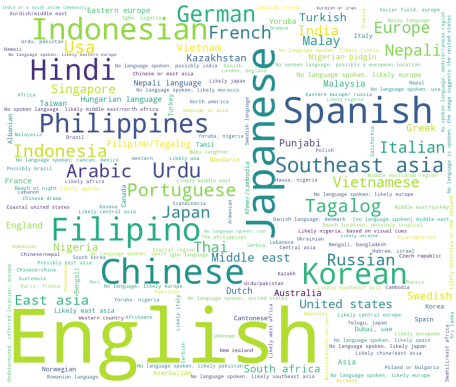


Figure 9: Language word-cloud in SIV-Bench.

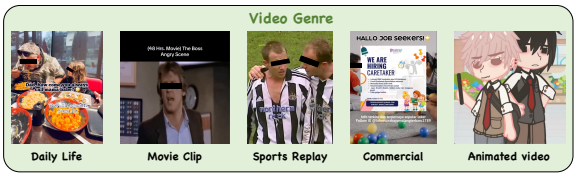


Figure 10: Examples illustrating the diverse video genres in SIV-Bench, including (from left to right) Daily Life recordings, Movie Clips, Sports Replays, Commercials, and Animated Videos, ensuring a broad range of social interaction contexts.

Figure 11: Examples of diverse video presentation styles featured in SIV-Bench, including (from left to right) Third-Person views, First-Person perspectives, Phone Call interfaces, and Solo Multi-Role Sketches.

B QA Details

B.1 Automatically QA Generation

The prompt shown in Figure 12 guides the Gemini in generating diverse and challenging Question-Answer (QA) pairs that form the foundation of SIV-Bench. Its design ensures these QAs effectively cover our core evaluation dimensions.

To create the multiple-choice options for each QA pair, we utilize a dedicated prompt for distractor generation, as detailed in Figure 13. This prompt instructs GPT to produce four unique distractor options based on the given question and its corresponding correct answer. Crucially, the prompt emphasizes that these distractors must be reasonable and clearly distinguishable from the correct answer, while also maintaining consistency with the correct answer in terms of sentence length, language style, and grammar. Furthermore, the prompt specifies a JSON output format for the list of distractors to ensure seamless programmatic integration. This structured approach to distractor generation is essential for constructing high-quality, fair multiple-choice questions that minimize biases arising from option formatting and robustly test model comprehension.

To classify each QA pair in SIV-Bench, we use a GPT model guided by a structured prompt (see Figure 14). The prompt defines our three main assessment dimensions—SSU, SSR, and SDP—and their sub-tasks. The model analyzes each QA and outputs its corresponding sub-task, ensuring consistent and precise categorization across the benchmark.

To classify each QA pair in SIV-Bench, we use a GPT model guided by a structured prompt (see Figure 14). The prompt defines our three main assessment dimensions—SSU, SSR, and SDP—and their sub-tasks. The model analyzes each QA and outputs its corresponding sub-task, ensuring consistent and precise categorization across the benchmark.

B.2 Human Annotation

Annotator Information. For the human annotation tasks integral to SIV-Bench, we recruit a team of 20 annotators. This cohort consists of 12 female and 8 male individuals, with an average age of 27 years. All annotators are well-educated and demonstrate proficiency in English, which is the primary language for instructions and annotations. To ensure fair compensation for their detailed work, annotators are remunerated at an hourly rate of $4 USD.

You will be given a video depicting human social relationships. Your task is to generate **question-answer (QA) pairs** to test the model's social reasoning ability. Each question must be clear and focused, addressing only one aspect at a time.

Question Requirements
- The questions should be challenging, requiring complex reasoning, an understanding of social norms, or real-world knowledge. Avoid surface-level observations.
- Generate **8-10 QA pairs**, distributed as follows:
- Descriptive: Focus on different aspects of the scene, but you don't need to cover all of them. You can choose what you find most important:
- Verbal characteristics (e.g., tone, pitch, speech style, etc.)
- Non-verbal characteristics (e.g., facial expressions, body language, etc.)
- Environmental features (e.g., location, setting, background details, etc.)
- Human interaction elements (e.g., emotion, feelings, attitude, reactions, etc.)
The question should be specific and unambiguous. Avoid questions like "describe the non-verbal characteristics of the video."
- Explanatory: Questions that explore the reasons behind observed behaviors.
- Predictive: A question that asks what is likely to happen next.
- Counterfactual: A question that explores hypothetical changes. For example, consider how the interaction might change if the individuals in the video had a stronger or weaker relationship, or if their relationship were of a different nature."

Answer Requirements
- Give a confident answer. Do not use words like "could" or "might." Provide only the essential information without additional explanations. Avoid making assumptions about relationships or identities in the QA. For example, DO NOT use descriptions like "the father and son in the video."
The output should be formatted in **JSON** as follows:
```json
{ "qa_pairs": [ { "category": "Descriptive", "question": "...", "answer": "..." }, { "category": "Explanatory", "question": "...", "answer": "..." }, { "category": "Predictive", "question": "...", "answer": "..." }, { "category": "Counterfactual", "question": "...", "answer": "..." } ] }
```

Figure 12: The prompt used to guide Gemini for the initial generation of question-answer pairs.

Given the following question and answer, generate four distractors that are reasonable and distinguishable from the correct answer.
The sentence length, language style and grammer should be consistent with the answer.\n\n"
f"Question: {qa_pair.get("question", "none")}\n"
f"Answer: {qa_pair.get("answer", "none")}\n\n"
"Provide the output in the following JSON format: \n"
"{\"distractors\": [\"<distractor 1>\", \"<distractor 2>\", \"<distractor 3>\", \"<distractor 4>\"]}"

Figure 13: The prompt used to generate distractors for each QA pair.

Annotator Guidelines. To maintain consistency and high quality in human annotation, particularly for the creation of new challenging Question-Answers (QAs) for Subset 2, annotators are provided with a comprehensive set of guidelines, as illustrated in Figure 15. These instructions detail the process for responding to existing multiple-choice questions and, crucially, guide annotators in formulating one new challenging question per video. For this QA creation task, annotators are directed to ensure their questions primarily test one of our three core assessment dimensions: Social Scene Understanding (SSU), Social State Reasoning (SSR), or Social Dynamics Prediction (SDP). The guidelines, including example questions, also emphasize that new QAs must be demanding, clearly worded, and require a deep understanding of the video content rather than superficial observation.

Annotation Process. All human annotation tasks for SIV-Bench, including the review of existing Question-Answers (QAs) and the generation of new QAs, are conducted using a custom-designed web interface. Figure 16 provides a representative example of this interface, which allows annotators to view the social interaction video, respond to provided multiple-choice questions, and submit their own newly authored questions and answers based on the video content and provided guidelines. This platform ensures a standardized environment for all human annotation contributions.

Quality Control and Inter-Annotator Agreement. To quantitatively address label quality, we compute the Inter-Annotator Agreement (IAA) using Fleiss' Kappa. For the initial set of approximately 27,000 QA pairs generated by our three LLM experts, the resulting Fleiss' Kappa is 0.52, indicating "moderate" agreement. For the questions that the LLMs disagreed, we measure the

For the understanding and reasoning of a social interaction video, it needs to be examined from the following three aspects of ability:
SSU (Social Scene Understanding): Action Recognition, Facial Expression Recognition, Human Attribute Identification, Environment Perception;
SSR (Social State Reasoning): Relation Inference, Emotion Inference, Intent Inference, Attitude Inference;
SDP (Social Dynamics Prediction): Factual Prediction, Counterfactual Prediction;
Now you need to analyze and categorize a multiple-choice question, output which level of test it belongs to and what aspects it specifically tests.
Here is the content:

Question: "{question}"
Options: "{options}"

Your output should be in the following format:
SSR: Relation Inference.

No other text or explanations are permitted.

Figure 14: Prompt used to instruct LLM for the final classification of Question-Answer pairs into one of the 10 fine-grained sub-tasks under SSU, SSR and SDP.

agreement between our human annotators, yielding a Fleiss' Kappa of 0.68. According to established standards (Landis and Koch, 1977), this value represents "substantial" agreement. Achieving this on the most difficult portion of our dataset confirms that humans can establish a reliable ground truth and justifies our methodology of retaining only those questions on which all assigned annotators unanimously agreed for the final benchmark.

B.3 QA statistics

To ensure the quality and balance of our SIV-Bench Question-Answer (QA) pairs, we conducted a statistical analysis of their structural properties, as summarized in Figure 17. This analysis examines several aspects: the average word count per multiple-choice option (Figure 17a) is relatively consistent across options A through E, minimizing length-based cues. The distribution of correct answers (Figure 17b) is fairly uniform across the five options, preventing positional bias. Furthermore, the overall question length (Figure 17c) peaks at around 11 words but shows a broad range, indicating variability in question complexity. Finally, an analysis of the first word in questions (Figure 17d) reveals a diverse set of interrogative types, led by 'what,' 'why,' and 'which,' reflecting a variety of reasoning challenges posed to the models.

B.4 Quality and Diversity Analysis

To further validate the quality of SIV-Bench, we conduct two analyses to ensure our benchmark encourages genuine comprehension over fitting to superficial patterns.

B.4.1 Analysis of Template Pattern Exploitation

SIV-Bench organizes tasks into SSU, SSR, and SDP to evaluate distinct dimensions of social cognition. To directly test whether models exploit surface-level cues, we analyze the properties of questions that a representative model (Gemini-2.0-Flash) answers correctly versus incorrectly. As shown in Table 5, while statistically significant differences in length and word count exist, the small effect sizes (Cohen's d) indicate no meaningful structural separation between the two groups. Furthermore, Table 6 shows a high cosine similarity between the embeddings of the correct and incorrect QA sets, along with similar internal distributions (intra-similarity and variance). This close alignment suggests that model performance is unlikely to rely on superficial statistical or semantic patterns.

B.4.2 Linguistic Diversity Analysis

The majority of questions in SIV-Bench are generated by large language models or written by human

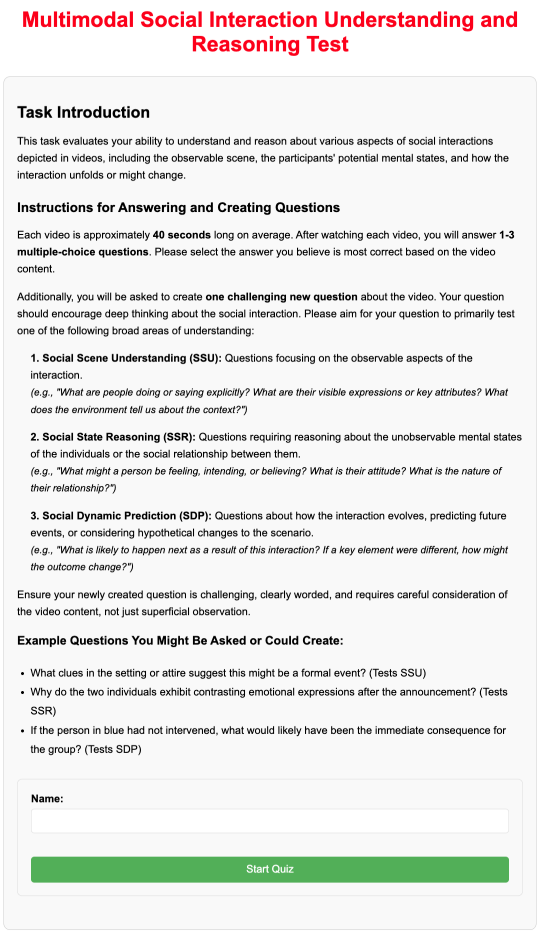


Figure 15: Screenshot of the guidelines provided to human annotators, detailing the tasks of answering multiple-choice questions and creating new challenging questions about Social Scene Understanding (SSU), Social State Reasoning (SSR), and Social Dynamics Prediction (SDP).

Table 5: Length and word count comparison between correctly and incorrectly answered questions.

Metric	Mean Length (chars)	Word Count
Correct	251.28 ± 93.88	42.13 ± 15.50
Wrong	228.96 ± 94.49	38.21 ± 15.78
t-stat	9.05	9.54
p-value	2.36e-19	2.63e-21
Cohen's d	0.2373	0.2516

annotators, rather than using rigid templates. To objectively measure our benchmark's linguistic diversity, we compared it against several prominent video QA benchmarks using two standard metrics: **Mean Semantic Distance** (average pairwise cosine distance of sentence embeddings) and **Vector Variance** (average variance across embedding dimensions). As shown in Table 7, SIV-Bench

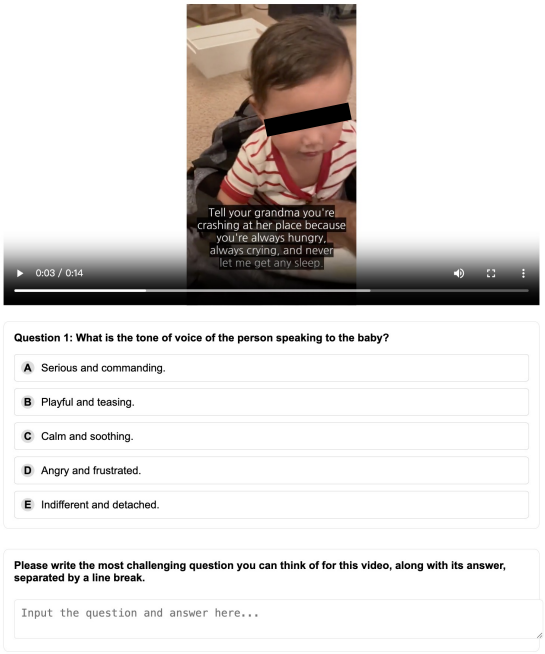


Figure 16: Example of the web-based interface used by human annotators for watching videos, answering provided multiple-choice questions, and authoring new Question-Answer pairs for SIV-Bench.

Table 6: Semantic similarity analysis between correctly and incorrectly answered question sets.

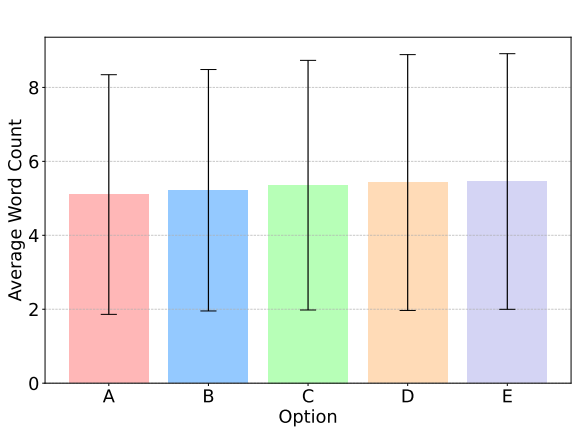
Metric	Value
Cosine Similarity (TF-IDF)	0.9765
Cosine Similarity (SentenceTransformer)	0.9759
Correct Intra-similarity	0.1689
Wrong Intra-similarity	0.1711
Correct Embedding Variance	0.002164
Wrong Embedding Variance	0.002157

exhibits high semantic diversity, ranking among the top benchmarks. These results support that SIV-Bench's questions are varied and not limited to shallow templates, thereby promoting genuine semantic understanding.

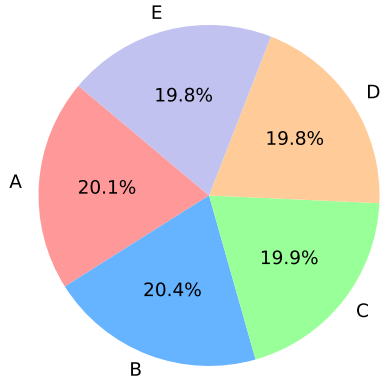
C Experimental Details

C.1 Settings Details

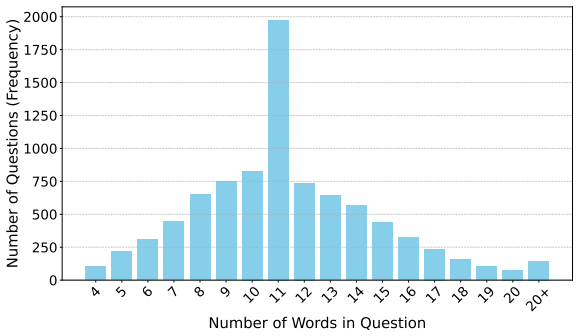
This section provides further details on our experimental setup for evaluating MLLMs on SIV-Bench. Figure 18 displays the standardized prompt templates employed for model evaluations, with distinct versions tailored to models based on their input capabilities. For models that process sequences of images, the "PROMPT for frames input" (Figure 18, Top) informs the MLLM that it will receive



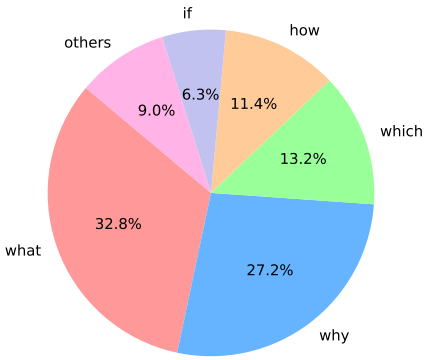
(a) Word count per option



(b) Correct answer distribution



(c) Question word count



(d) First word in questions

Figure 17: Statistical analysis of SIV-Bench question–answer (QA) pairs. (a) Average word count consistency across answer options, (b) distribution of correct answers among options, (c) question length distribution by word count, and (d) frequency of common first words in questions.

Table 7: Semantic diversity comparison across video QA benchmarks.

Benchmark	**Mean Dist.** ↑	**Variance** ↑
SIV-Bench	0.8321	0.0022
Video-Bench	**0.8811**	**0.0023**
Perception_Test	0.7677	0.0020
VideoVista	0.7604	0.0020
Social-IQ 2.0	0.7260	0.0019
MVBench	0.7124	0.0019
EgoSchema	0.5411	0.0014

a set of uniformly sampled frames from a video in chronological order. For models capable of direct video processing, the "PROMPT for videos input" (Figure 18, Bottom) is used. Both prompts clearly instruct the MLLM on its role, the task of answering multiple-choice questions based on the provided visual input, the expected JSON-like format for organizing answers (providing the exact text of the chosen option for each question), and a strict directive to avoid any extraneous text such as explanations or conversational remarks. This standardized, yet input-adaptive, prompting approach ensures consistency in task presentation across different model architectures.

For all evaluations, the specific inference parameters used for each model—such as temperature, top-p, or maximum new tokens—are adopted from their default configurations as provided within the VLMEvalKit (Duan et al., 2024) framework. This adherence to default settings aims to reflect the out-of-the-box capabilities of these models and ensure fair comparability. The experiments are conducted on two primary compute clusters. Cluster 1, utilized for evaluating the largest open-source models (Qwen2.5-VL-72B-Instruct and InternVL-78B), is equipped with an AMD EPYC 7642 48-Core Processor and 4x NVIDIA A100 GPUs. The total

runtime for the reported experiments on this cluster is approximately 3 days. Cluster 2, used for the remaining models, consists of an Intel(R) Xeon(R) Platinum 8369B CPU @ 2.90GHz and 8x NVIDIA RTX 3090 GPUs. The cumulative runtime for experiments on this cluster is approximately 2 days. It should be noted that the overall research project, including preliminary testing on earlier dataset versions and exploratory experiments not included in the final results, involved a greater amount of compute time than the specific durations reported for the final benchmark evaluations.

C.2 Failure Cases

This section presents several illustrative failure cases. We focus on examples from Gemini-2.0-Flash, a strong closed-source model, to highlight that even advanced MLLMs can fail on nuanced aspects of social perception, prediction, and reasoning. These examples are categorized by our primary assessment dimensions: SSU, SSR, and SDP, and are intended to offer concrete instances for future research and model development.

Figure 19 illustrates instances where Gemini-2.0-Flash fails on SSU tasks, which require accurate perception of explicit visual elements. **(a) In Action Recognition**, the model incorrectly identifies the man's gesture as "He crosses his arms tightly" instead of the correct "He raises one eyebrow slightly", missing a subtle but distinct facial action. **(b) For Environment Perception**, when asked about the weather, the model failed to capture the details of the characters in the scene wearing thick scarves and down jackets to infer that the correct answer was "cold", but instead wrongly chose "wet". **(c) In Facial Expression Recognition**, the model describes the expression as "A mischievous smile" rather than the correct "A stoic glare", misinterpreting the nuanced facial expression display. **(d) For Human Attribute Identification**, concerning the child's clothing, the model selects "A dress" instead of the correct "A set of pajamas", failing to correctly identify common apparel.

Figure 20 presents failure cases of Gemini-2.0-Flash on SSR tasks, which involve inferring unobservable mental states and relationships. **(e) In Intent Inference**, when a woman says "do you understand?" to a boy who bullies her son in an angry tone, it is to teach him a lesson and warn him not to bully her son again, not for "discourage any defiance", because in fact no child much younger than her can form defiance against her. **(f) For Emotion Inference**, this employee is happy instead of scared after leaving because he successfully deceives the boss into giving him a vacation. **(g) In Attitude Inference**, the coworker is dissatisfied and disappointed with the cashier's nervousness, panic and even physical reactions when seeing female customers. This could also be seen from his subsequent warning to the cashier not to do so anymore. **(h) For Relation Inference**, we present case studies on the failure patterns of the four common models listed in the main text in this task.

Figure 21 highlights errors made by Gemini-2.0-Flash in SDP tasks, which require predicting future events or reasoning about hypothetical scenarios. **(i) In Factual Prediction**, when asked if the person in the black shirt would be satisfied with the workers' work, since the two of them have already reached an agreement with smiles at the end of the video, it could be inferred that the answer was "yes", but the model chooses another answer. **(j) For Counterfactual Prediction**, the video shows the dance interaction between a mother and her son. The question raised is what would happen if one of them had more dance experience. This can be inferred from the positive and relaxed interaction between the two in the video. The most likely answer is "The more experienced dancer leads and adapts movements". For example, a son leads his mother to learn dancing happily, rather than "The less experienced dancer hesitates and struggles to keep up". They have a good relationship, and the probability of negative performance like "hesitates" and "struggles" is lower.

C.3 Analysis of Chain-of-Thought Prompting

To investigate the impact of explicit reasoning on model performance, we conducted a preliminary experiment using a Chain-of-Thought (CoT) prompting strategy. We prepended the instruction *"Let's think step by step. First, output your reasoning process, and then output the final answer."* to our standard evaluation prompt. The overall accuracy on the `origin` videos, with and without CoT, is presented in Table 8.

The results indicate that applying a generic CoT prompt did not yield significant performance improvements for most models. For several smaller open-source models (e.g., mPLUG-Owl3, LLaVA-Video), it resulted in a notable performance decrease. We observed that this is often because these models struggle to consistently adhere to the more complex two-stage output format (i.e., providing

PROMPT for frames input

You are an AI assistant responsible for answering questions about videos.

You will be provided with {} separate frames uniformly sampled from a video, \
the frames are provided in chronological order of the video.
Please analyze these images and provide the answers to the \
following multiple-choice questions about the video content.
If multiple questions are provided (with indices Q1, Q2, Q3, ...), \
you should organize your answers in the following json format:
1. [Exact text of the chosen option for question 1],
2. [Exact text of the chosen option for question 2],
...

Do NOT add any explanations, introductions, or concluding remarks.

PROMPT for videos input

You are evaluating a video based on the multiple-choice questions provided below.,
For each numbered question, select the best answer from the options listed.,
Your response MUST strictly follow this format:,
1. [Exact text of the chosen option for question 1],
2. [Exact text of the chosen option for question 2],
... and so on for all questions.

Do NOT add any explanations, introductions, or concluding remarks.,
--- QUESTIONS ---
…

Figure 18: Standardized prompt templates used for evaluating MLLMs on SIV-Bench. Separate prompts are shown for models that process (Top) uniformly sampled frames and (Bottom) direct video input.

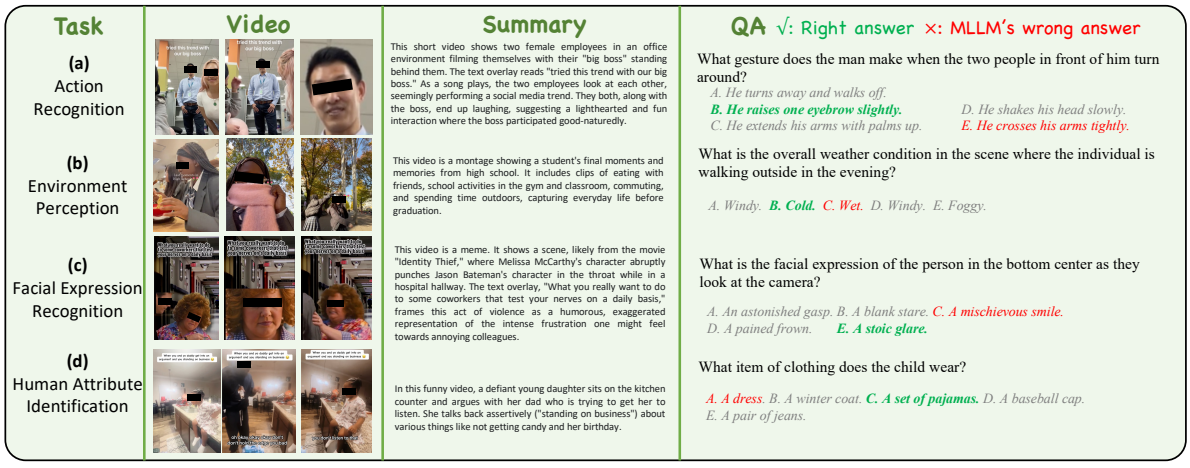


Figure 19: Examples of failure cases in Social Scene Understanding (SSU) tasks, including errors in Action Recognition, Environment Perception, Facial Expression Recognition, and Human Attribute Identification.

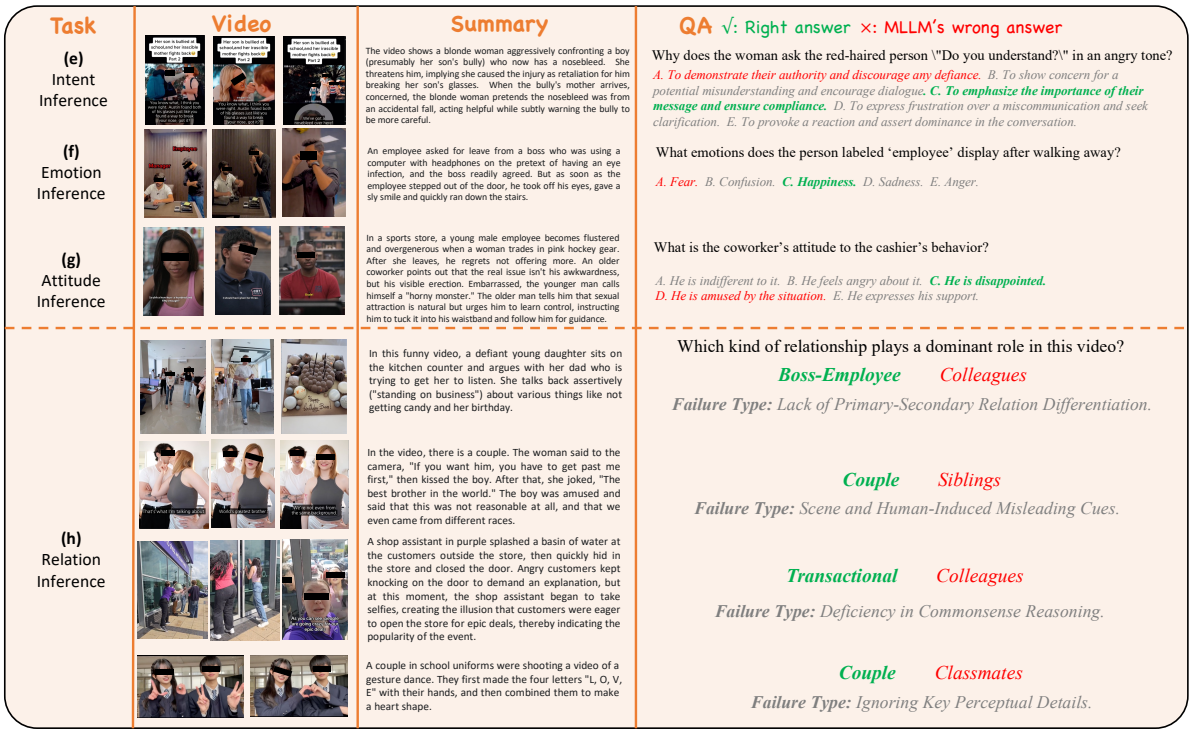


Figure 20: Examples of failure cases in Social State Reasoning (SSR) tasks, highlighting difficulties in Intent Inference, Emotion Inference, Attitude Inference, and Relation Inference.

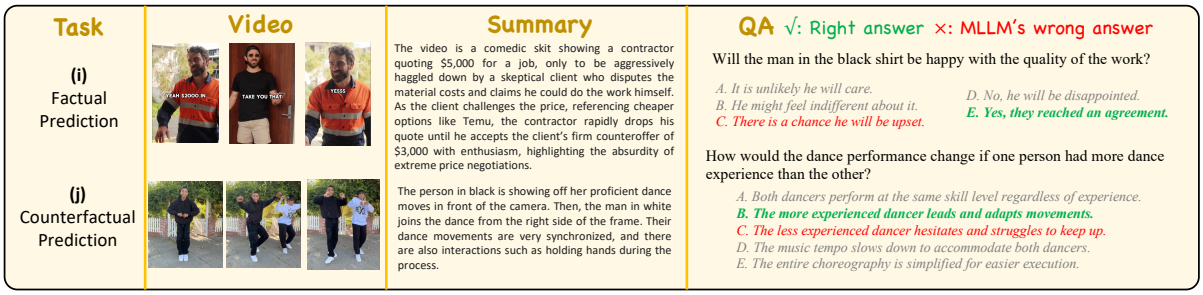


Figure 21: Examples of failure cases in Social Dynamics Prediction (SDP) tasks, covering both Factual Prediction and Counterfactual Prediction.

reasoning before the final answer in the required format), leading to failures in our answer parsing logic.

The primary goal of SIV-Bench is to establish a fair, consistent, and reproducible evaluation of baseline model capabilities. Our standardized prompting strategy, which aligns with widely used toolkits like VLMEvalKit, ensures this fairness. While techniques like CoT are powerful for eliciting maximum performance from certain capable models (e.g., Gemini-2.5-Pro), introducing them as a default can create a confounding variable. Such a setup might shift the evaluation from testing inherent social reasoning to testing complex instruction-following abilities. Therefore, our main experiments use a direct-answering prompt to maintain a level playing field. Nonetheless, these findings suggest that developing more specialized reasoning methods tailored to social intelligence is a valuable direction for future work that builds upon this benchmark.

Table 8: Comparison of Overall Accuracy (%) on the `origin` videos with and without Chain-of-Thought (CoT) prompting.

Model	Acc (Origin)	Acc (CoT)
mPLUG-Owl3	42.06	41.79
LLaVA-OneVision	41.97	41.22
LLaVA-Video	41.09	40.13
Qwen2.5-VL-7B-Instruct	44.02	44.76
InternVL-8B	45.82	44.99
Qwen2.5-VL-72B-Instruct	58.80	59.13
InternVL-78B	55.46	55.25
o4-mini	55.68	55.79
GPT-4o	58.02	57.91
Gemini-2.0-Flash	56.40	56.53
Gemini-2.5-Flash	57.87	57.71
Gemini-2.5-Pro	61.65	61.77

C.4 Statistical Significance Analysis

To validate the reliability of our comparative claims, we conducted McNemar's tests on the full dataset ($N = 5,455$). This paired non-parametric test is appropriate for comparing the performance of two classifiers on the same dataset.

Model Ranking. We verified the leadership of our SOTA model. The performance difference between the top-performing **Gemini-2.5-Pro** (61.65%) and the second-best model **Qwen2.5-VL-72B** (58.80%) is highly statistically significant ($p < 0.001$), confirming the robustness of the leaderboard rankings.

Task Difficulty. We confirmed that the performance stratifications across our three core dimensions are not due to chance. For Gemini-2.5-Pro, the performance gaps between **SSU** (85.07%) and **SDP** (60.45%), as well as between **SDP** and **SSR** (54.30%), are all highly statistically significant ($p < 0.001$).

Subtitle Influence. We performed significance testing on the subtitle conditions (Table 4). Our analysis reveals that the minor *Overall* improvement from adding subtitles ('+sub') is not statistically significant ($p = 0.12$). However, the impact is task-dependent: the negative impact of removing text ('-sub') is statistically significant for the **SSR** task ($p < 0.05$), and the benefit of added subtitles ('+sub') is significant for the **SDP** task ($p < 0.05$).

D SIV-Bench-Hard Details

To rigorously evaluate the upper limits of current MLLMs, establish a robust human baseline, and probe the reasoning processes beyond simple accuracy, we curated and analyzed a challenging subset of our dataset, termed **SIV-Bench-Hard**. This section details the setup, human performance, and a multi-dimensional analysis of model reasoning quality on this subset.

D.1 Experimental Setup

Dataset Curation. We selected a subset of 200 questions from the human-generated portion of SIV-Bench. These questions were specifically chosen for their complexity and reliance on deep social understanding, filtering out items that could be solved via superficial visual cues.

Task Definition. Unlike the standard multiple-choice evaluation, this study required both human annotators and MLLMs to provide: (1) the selected answer option, and (2) a free-text *reasoning explanation* justifying their choice. The prompt is shown in Figure 22 and 23. This allows for a deeper examination of the cognitive process.

Participants. We recruited 3 independent human annotators to perform this task to establish a human baseline. We evaluated a suite of state-of-the-art MLLMs, including Gemini-3-Pro, GPT-5.1, Gemini-2.5-Pro, Gemini-2.5-Flash, Qwen2.5-VL-7B, and GPT-4o-mini.

```
Please watch the video and answer the following multiple-choice question.

Question: {question}

Options:
{options_text}

Please provide your answer strictly in the following format:

Answer: [Your chosen option letter, e.g., A/B/C/D/E]
Explanation: [Brief explanation in less than 50 words]

Important:
1. You must choose one option from the given choices
2. Keep your explanation concise and clear
3. Strictly follow the format above
```

Figure 22: Model prompt for SIV-Bench-Hard.

D.2 Quantitative Analysis of Reasoning

To quantify the divergence observed in the qualitative scores, we performed an embedding-based analysis. We encoded all reasoning texts (both human and model) using the `paraphrase-multilingual-MiniLM-L12-v2` model.

Distinguishability. We trained a Random Forest classifier to distinguish between human and model explanations based on their embeddings. The classifier achieved an accuracy of **92.17%** (compared to a 50% random baseline). This high classification accuracy indicates that the latent semantic features of model reasoning are fundamentally distinct from those of humans.

Statistical Significance. We further validated this difference using Mann-Whitney U tests across the embedding dimensions. The tests confirmed that the distributions of human and model embeddings are significantly different ($p < 0.05$) on 8 out of 10 principal dimensions. Collectively, these results provide quantitative evidence that current MLLMs, despite their linguistic fluency, employ reasoning processes that are statistically distinguishable from human social cognition.

The SIV-Bench-Hard subset, along with the human reasoning annotations and analysis code, will be released to facilitate future research into bridging this gap.

```
You are an expert evaluator assessing the quality of AI model reasoning compared to human reasoning.

**Task Context:**
Question: {question}
Options: {options}

**Human References (3 human reasonings):**
Human 1: {human1_reason}
Human 2: {human2_reason}
Human 3: {human3_reason}

**Model Response to Evaluate:**
Model: {model_name}
Model's Answer: {model_answer}
Model's Reasoning: {model_reason}

**Evaluation Instructions:**
Please evaluate the model's reasoning quality across the following dimensions. Remember, we assume human-centered AI development, so human reasoning processes are the gold standard to approximate, even if a human's choice is incorrect.

Rate each dimension on a scale of 1-5:
- 1: Very Poor
- 2: Poor
- 3: Fair
- 4: Good
- 5: Excellent

**Evaluation Dimensions:**

1. **Relevance ()**: Does the reasoning address the key aspects of the question and video content?
2. **Alignment with Human Reasoning ()**: How similar is the model's reasoning process to the human references? Does it consider similar factors and perspectives?
3. **Logical Coherence ()**: Is the reasoning logically sound and internally consistent?
4. **Depth of Analysis ()**: Does the reasoning show deep understanding of social relations, emotions, and context?
5. **Conciseness ()**: Is the reasoning clear and concise without unnecessary verbosity?

**Response Format (JSON only):**
{{
  "relevance": {{
    "score": <1-5>,
    "justification": "<brief explanation>"
  }},
  "alignment_with_human": {{
    "score": <1-5>,
    "justification": "<brief explanation>"
  }},
  "logical_coherence": {{
    "score": <1-5>,
    "justification": "<brief explanation>"
  }},
  "depth_of_analysis": {{
    "score": <1-5>,
    "justification": "<brief explanation>"
  }},
  "conciseness": {{
    "score": <1-5>,
    "justification": "<brief explanation>"
  }},
  "overall_score": <average of 5 dimensions>,
  "overall_comment": "<2-3 sentences summarizing the model's reasoning quality>"
}}

Respond ONLY with valid JSON. Do not include any text before or after the JSON.
```

Figure 23: LLM-judge prompt for scoring explanations in SIV-Bench-Hard.

D.3 Correlation Between Answer Correctness and Reasoning Fidelity

To rigorously verify that model performance is grounded in genuine social reasoning rather than superficial visual shortcuts (e.g., background cues or object co-occurrence), we analyze the relationship between the correctness of a model's answer and the semantic quality of its reasoning trace. We hypothesize that if models were merely "guessing" via shortcuts, their generated explanations would lack alignment with human cognitive processes even when they fortuitously select the correct option.

We perform an embedding-based analysis on the *SIV-Bench-Hard* subset using the `paraphrase-multilingual-MiniLM-L12-v2` model. For each of the six evaluated models, we calculate the cosine similarity between the model's generated reasoning and the human expert ground truth. The results are stratified based on whether the model answers the multiple-choice question correctly or incorrectly.

Table 9: Quantitative comparison of reasoning similarity to human ground truth between correct and incorrect responses. The **Gap (Δ)** represents the increase in similarity when the model answers correctly. Significance is calculated using the Mann-Whitney U Test.

Model	Similarity (Correct)	Similarity (Incorrect)	Gap (Δ)	Significance (p-value)
GPT-5.1	0.572	0.506	+0.066	$<$ **0.001** (***)
Qwen2.5-VL-7B	0.535	0.474	+0.061	$<$ **0.01** (**)
GPT-4o-mini	0.571	0.513	+0.058	$<$ **0.01** (**)
Gemini-2.5-Pro	0.541	0.498	+0.043	$<$ **0.05** (*)
Gemini-2.5-Flash	0.537	0.504	+0.033	$<$ **0.05** (*)
Gemini-3-Pro	0.527	0.494	+0.033	0.067 (n.s.)

As illustrated in Figure 24 and summarized in Table 9, we observe a statistically significant positive gap in similarity scores across 5 out of the 6 models. Notably, GPT-5.1 demonstrates the strongest effect, with a similarity gap of 0.066 ($p < 0.001$).

These findings provide empirical evidence that correctness in SIV-Bench is strongly correlated with human-like social reasoning. The significant degradation in reasoning alignment during failure cases suggests that models do not rely on "guessing" via superficial cues; rather, successful performance necessitates a cognitive process that mirrors human social understanding.

E Broader Impact

SIV-Bench is designed to foster positive advancements in artificial social intelligence, potentially leading to more empathetic, context-aware, and collaborative AI systems for beneficial applications such as assistive technologies, improved human-AI teaming, and richer content understanding. However, enhancing AI's grasp of social dynamics also presents risks. These capabilities could be misused for sophisticated manipulation, disinformation, or invasive surveillance, and unaddressed biases in data could be amplified, leading to inequitable outcomes. We offer SIV-Bench as a research tool to transparently assess MLLM capabilities and limitations in the social domain, thereby encouraging the community to proactively consider these ethical challenges and develop robust safeguards alongside continued innovation in social AI.

F Use of Ai Assistants

The LLM was utilized exclusively to aid and polish the writing, including for tasks such as improving grammar and clarity, refining sentence structure, and ensuring stylistic consistency

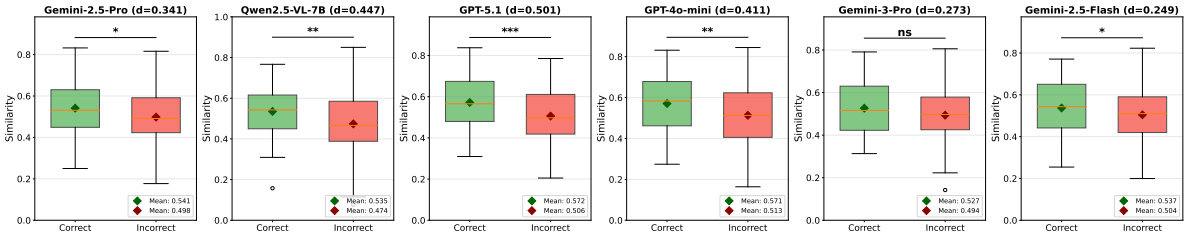


Figure 24: Box plots illustrating the distribution of semantic similarity scores between model reasoning and human ground truth, stratified by answer correctness. We observe a consistent trend where reasoning traces for correct answers exhibit significantly higher alignment with human social cognition compared to incorrect answers. Statistical significance is denoted by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).