

SMMU: Benchmarking Social Intelligence of Multimodal Large Language Models

Gordon Chen¹ Ziqi Huang¹ Ailing Zeng² Chuhan Mei¹ Kairui Hu¹

Xin Tong² Ziwei Liu¹

¹Nanyang Technological University ²Anuttacon

Project Page: <https://gordonchen19.github.io/SMMU/>

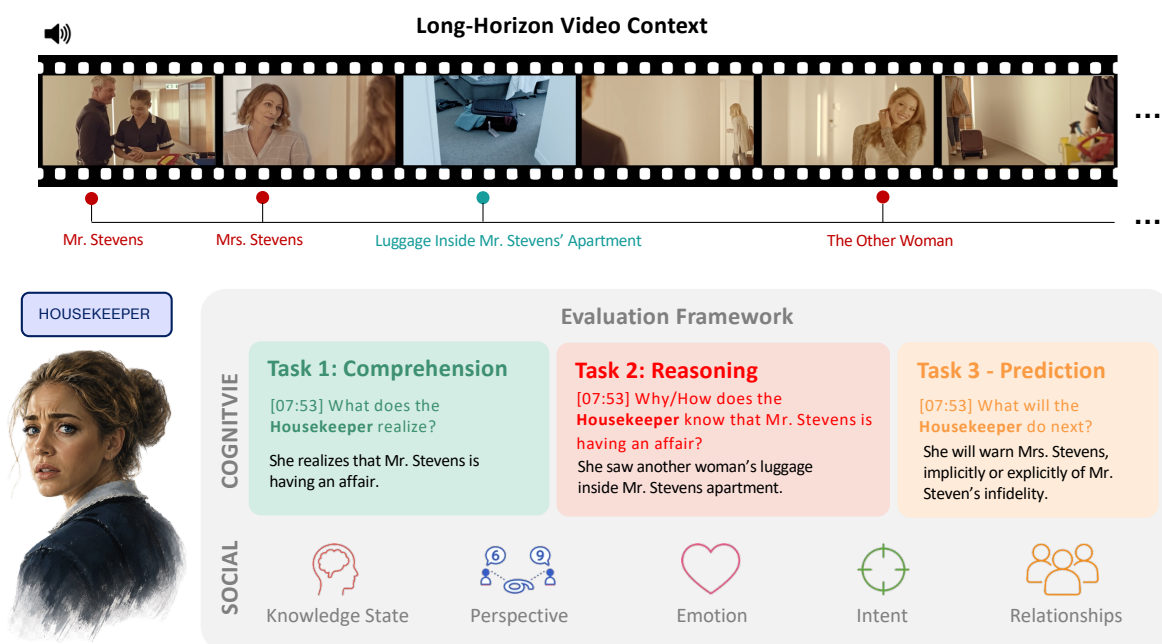


Figure 1: *SMMU* systematically evaluates Social Intelligence across 5 major social dimensions and across three cognitive tasks: (1) *Comprehension*; (2) *Reasoning*; and (3) *Prediction*.

Abstract

Recent Multimodal Large Language Models (MLLMs) have made rapid progress in video understanding, demonstrating strong performance across a wide range of perceptual and reasoning tasks. However, their *Social Intelligence* remains underexplored and difficult to evaluate. Existing benchmarks for social understanding suffer from two major limitations: they primarily evaluate short-horizon interactions and focus largely on explicit perceptual cues, such as facial expressions and speech sentiment. However, real-world social understanding often requires inferring implicit emotions, intentions, beliefs, and interpersonal dynamics from extended contextual interactions rather than isolated expressions or utterances. Consequently, robust evaluation of social intelligence requires long-horizon reasoning over complex social contexts. To address this gap,

we introduce *SMMU*, a **Massive Multimodal Social Understanding** benchmark. *SMMU* comprises 264 long-form audiovisual scenes and 1519 question-answer pairs, to systematically evaluate each model across five social dimensions (Emotions, Intent, Knowledge State, Perspective, Relationships) and three cognitive tasks (Comprehension, Reasoning, Prediction). *SMMU* is designed to require models to integrate contextual evidence across extended interactions in order to interpret, reason about, and anticipate complex human behavior. Evaluating a suite of state-of-the-art MLLMs, we find that even the strongest models achieve only 0.43 Social Understanding Score compared to the average human baseline of 0.85, and 0.33 Social Agency Score compared to the human baseline of 0.78, revealing a substantial gap between current MLLMs and human social intelligence.

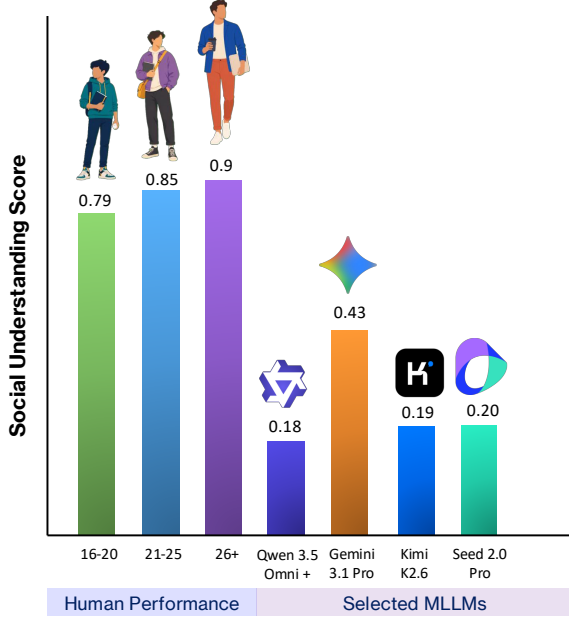


Figure 2: **Social Understanding Score across humans and selected MLLMs.** Human performance on SMMU increases slowly across age groups, while all evaluated MLLMs remain substantially below human levels, revealing a large gap in social understanding despite strong general multimodal capabilities.

1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have led to substantial progress in video understanding and are increasingly being deployed in settings that demand long context interactions with people, as personal assistants, tutors, healthcare companions, and embodied agents. This raises an important question: do current MLLMs exhibit human-level social intelligence? i.e., Can MLLMs understand, reason and manage others in complex social interactions (Thorndike and Stein, 1937)? Despite headline performance on perceptual and reasoning benchmarks, the level of social intelligence of current MLLMs remains unclear and difficult to evaluate.

Existing benchmarks for social understanding (Kong et al., 2025; Hu et al., 2025; Lian et al., 2025) remain limited in two important ways. First, they predominantly evaluate models on short clips, emphasizing localized perception rather than the ability to reason over extended, multi-turn social interactions. Second, they focus primarily on explicit perceptual cues, such as facial expressions and speech sentiment, despite the fact that many social and mental states in real-world interactions are inherently implicit and contextual. For instance,

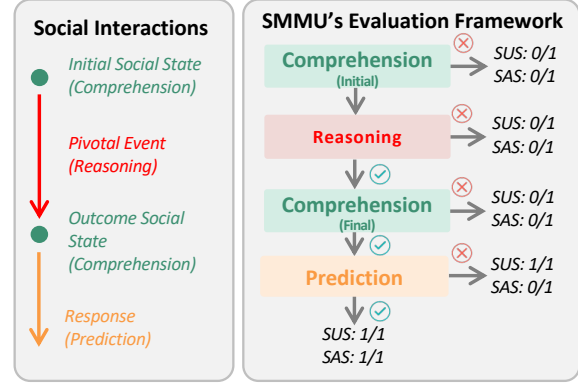


Figure 3: **SMMU's Joint Evaluation Framework.** We assess whether models maintain and appropriately update their social representations as events unfold over long-context videos. Models answer a chain of comprehension (C), reasoning (R), and prediction (P) questions. The Social Understanding Score (SUS) receives credit only when the required comprehension and reasoning questions are both correct; the Social Agency Score (SAS) additionally requires a correct prediction.

emotions and intentions are often concealed rather than directly expressed, while beliefs and perspectives frequently must be inferred from prior interactions and experiences. As a result, robust social understanding requires integrating contextual clues across long temporal horizons rather than relying solely on isolated expressions or utterances. These limitations underscore the need for more robust benchmarks to test social intelligence under long-horizon, multimodal, and reasoning-intensive settings. To address this gap, we introduce SMMU, a benchmark for social intelligence in video understanding. SMMU comprises 264 long-form audio-visual scenes and 1,519 multiple-choice QA pairs spanning five major social dimensions: Emotion, Intent, Perspective, Knowledge State, and Relationships, as well as three cognitive tasks: Comprehension, Reasoning, and Prediction. We designed highly challenging questions that cannot be solved through shallow perceptual recognition or short-context reasoning alone. Instead, models must integrate relevant evidence across extended interactions to understand the video.

Additionally, we introduce two evaluation metrics, the Social Understanding Score (SUS) and Social Agency Score (SAS), to assess whether models can consistently understand videos across multiple cognitive levels, including comprehension, reasoning, and prediction, thereby penalizing partial understanding and guess-based success. We bench-

Modality	Video Duration	Social Reasoning	Joint Evaluation
MME-Emotion	Short	Emotion	✗
AffectGPT	Short	Emotion	✗
Social-IQ	Short	✓	✗
SIV-Bench	Short	✓	✗
Social Genome	Short	✓	✗
SMMU (Ours)	Long	✓	✓

Table 1: Comparison of social understanding benchmarks. Unlike prior benchmarks that focus on independent MCQs, *SMMU* evaluates long-horizon social understanding through a staged Comprehension → Reasoning → Prediction protocol.

mark existing state-of-the-art audio-visual MLLMs across multiple model families including Gemini (DeepMind, 2026), Qwen (Alibaba, 2026), Kimi (Kimi, 2026), Seed (Bytedance, 2026). Our results reveal a substantial performance gap between current frontier models and humans, with the strongest model achieving a social understanding score of only 0.43 compared to the human expert baseline of 0.85. In particular, we find that current MLLMs significantly lack evidence-grounded causal reasoning for video understanding, limiting their ability to form stable social representations necessary for robust social intelligence.

2 Related Works

2.1 MLLMs for Video Understanding.

Recent advances in MLLMs have significantly expanded multimodal understanding capabilities from image-level perception to long-form video comprehension (Team et al., 2024, 2026). The emergence of omni-modal models further enables unified reasoning over visual, linguistic, and auditory signals, providing richer representations of complex real-world interactions (Google DeepMind, 2026; Xu et al., 2025a,b; Seed, 2026). Meanwhile, reasoning-focused approaches, including chain-of-thought prompting, supervised fine-tuning, and reinforcement learning, have been adapted to improve video reasoning abilities (Feng et al., 2025; Li et al., 2025; Yan et al., 2025). However, existing video understanding benchmarks primarily evaluate perceptual recognition or short-horizon reasoning, such as identifying events, actions, and objects. They provide limited assessment of whether models can integrate social cues to understand complex human interactions.

2.2 Social Intelligence Benchmarks.

Social intelligence requires understanding others’ behaviors, mental states, and social interactions beyond recognizing explicit emotional cues. While extensive efforts have studied emotional intelligence across text (Paech, 2023; Liu et al., 2024; Sabour et al., 2024), image (Yang et al., 2025), and short videos (Hu et al., 2025; Zhang et al., 2025; Lian et al., 2025), these capabilities represent only one aspect of broader social cognition. In contrast, evaluating whether models can infer latent intentions, beliefs, perspectives, and relationships underlying human interactions remains relatively underexplored. Early computational benchmarks focused on text-based social reasoning, with Social IQa evaluating whether models can infer social consequences, motivations, and reactions in everyday situations (Sap et al., 2019). Subsequent work investigated the ability of language models to perform theory-of-mind reasoning (Sap et al., 2022). With the emergence of multimodal models, Social-IQ (Zadeh et al., 2019), SIV-Bench (Kong et al., 2025) and Social Genome (Mathur et al., 2025) extended social intelligence evaluation to audiovisual interactions, requiring models to reason about human behavior from conversational and visual cues in brief social interactions. However, these benchmarks predominantly target short-horizon interactions, where the relevant cues -facial expressions, dialogue, vocal signals- are highly explicit within a single scene. Real-world social understanding, by contrast, typically depends on integrating subtle contextual evidence accumulated across extended interactions, meaning strong performance on existing benchmarks need not reflect robust long-horizon social understanding. *SMMU* addresses this gap.

3 SMMU

SMMU decomposes social understanding along two axes: social dimensions (what is inferred) and cognitive stages (how it is inferred) (Guilford, 1956). The 5 social dimensions are informed by Happé et al.’s taxonomy of social cognitive abilities (Happé et al., 2017), while the 3 cognitive dimensions is grounded in theories of mentalizing-social understanding as causal inference over hidden mental states that explain and predict behavior (Kliemann and Adolphs, 2018).

Plot Synopsis: A man (Frank) tries to impress his office crush by staging a fake purse robbery with the help of a co-worker, intending to rescue her and appear heroic. However, an actual robbery takes place in the building, and he mistakenly assumes the masked robber is his friend rather than a real criminal.





Frank

Knowledge State

[03:40.40 -> 03:57.59] Which of the following correctly describes what Frank knows or does not know at this point in the video?

A. The woman is in danger; Frank does not know that. [03:57.59]
 B. The woman is in danger; Frank is aware of that.
 C. Frank's friend is disguised as the robber; Frank does not know that.
 D. The woman is not actually in danger; Frank knows that. [03:40.40]

[03:57.59] Why is Frank unaware that the woman is in danger?

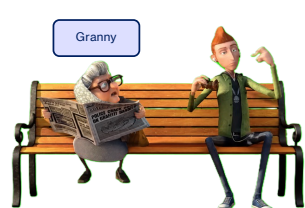
A. Because he does not know that the robber is not his friend.
 B. Because the robber looked inexperienced to him.
 C. Because he did not know that Frank's friend intends to harm the woman.
 D. Because Frank was distracted by a phone call from his ex-wife.

[03:57.59] Which of the following is most likely to occur next?

A. The robber will punch Frank.
 B. They will proceed with the planned robbery.
 C. Frank will flee out of fear of the actual robber.
 D. Frank will become immobilized by fear and do nothing.
 E. Frank will freeze and do nothing because he is scared.
 F. Frank will assist the robber in stealing the purse.

Plot Synopsis: An elderly woman grows increasingly frustrated with a young man sitting nearby who appears to be rudely stealing and eating her cookies, only to discover at the very end that she had actually been eating his cookies the entire time.





Granny

Emotion

[03:20.12 -> 03:46.78] What is the primary emotion that the granny is experiencing at this moment in the video?

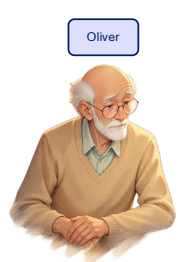
A. Confused
 B. Guilty [03:46.78]
 C. Frustrated [03:20.12]
 D. Relieved

[03:46.78] Which explanation best describes why the granny feels guilty?

A. She realized that the teenager had been trying to share his own snack with her.
 B. She realized that the teenager had bought her an unopened box of cookies as a kind gesture.
 C. She realized that the teenager had missed his train because their disagreement distracted him.
 D. She realized that the teenager had given up his cookies after seeing that she could not get any from the vending machine.

Plot Synopsis: In a bio-technologically advanced society where humans can be reborn into younger bodies without retaining their memories, an elderly man rejects the process, choosing instead to preserve the memories of his late wife.





Oliver

Intent

[03:41.02 -> 06:14.83] Which of the following correctly describes what Oliver's intent is at this point in the video?

A. He intends to sign the rebirth certificate.
 B. He does not intend to sign the rebirth certificate. [03:41.02 -> 06:14.83]
 C. He intends to postpone signing until he has spoken to Wendy again.
 D. He intends to ask the clinic to reverse Wendy's rebirth procedure.

[06:14.83] Why does Oliver not intend to sign the rebirth certificate in this scene, or how does he intend to achieve this?

A. He does not want to lose his memories.
 B. He suspects the clinic is hiding risks associated with the rebirth program.
 C. He does not want a new body.
 D. He believes that life has meaning only because it is finite.

[06:14.83] What is Oliver most likely to experience next?

A. Oliver will die from old age.
 B. Oliver will be reborn with no recollection of his past.
 C. Oliver will be reborn with memories of his past.
 D. Oliver will explain to Wendy about the life they shared.

Figure 4: **SMMU Question Samples.** SMMU structures evaluation as a chain of cognitive tasks: Comprehension (Green), Reasoning (Red), and Prediction (Yellow). Each question only has 1 defensible answer.

3.1 Social Dimension Taxonomy

We evaluate Social Intelligence across 5 core social dimensions.

Emotions. Whether models can infer a character's affective state from behavioral and contextual cues.

Intent. Whether models can infer the underlying goals or motivations driving a character's actions.

Perspective. Whether models can adopt a character's point of view on a situation.

Knowledge State. Whether models can track what a character knows or doesn't know.

Relationships. Whether models can infer interpersonal relationship dynamics such as likeness, trust and power balance.

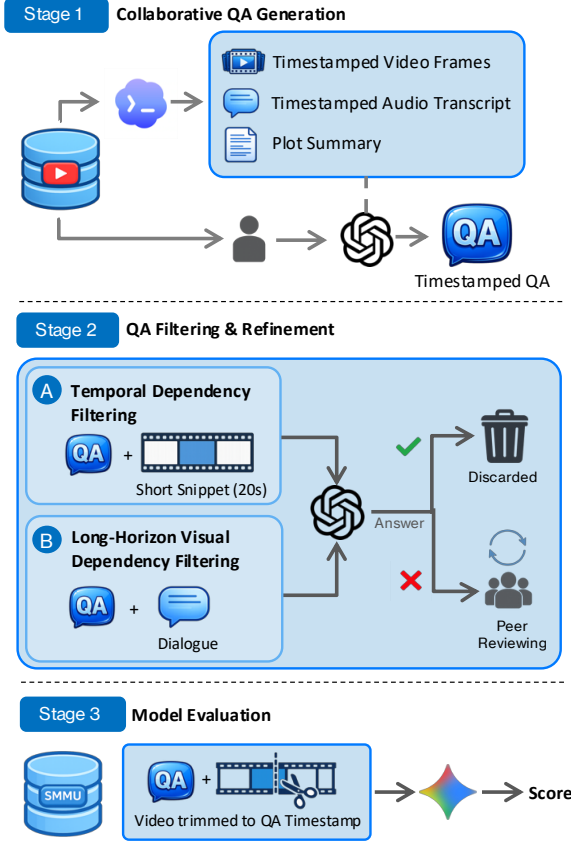


Figure 5: SMMU Annotation Pipeline.

3.2 Cognitive Dimension Taxonomy

We evaluate the model across 3 cognitive tasks, each demonstrating a different area of understanding, allowing us to distinguish between different cognitive failure modes.

Comprehension. Whether the model can maintain and update their social understanding. We ask two questions, one before and one after a pivotal event to test whether the model’s social representation is appropriately tracked and revised.

Reasoning. Whether the model can identify the underlying evidence or causal factors supporting a given social interpretation. The reasoning question is conditioned on the ground-truth answer to the corresponding comprehension question.

Prediction. Whether the model can anticipate plausible future actions, reactions, or outcomes.

3.3 Evaluation Metrics

SMMU evaluates both the *breadth* and *depth* of social understanding to provide a more interpretable measure of true social intelligence.

Evaluating Breadth of Understanding. We report mean accuracy across all questions, as well

as across each social dimension and cognitive task. This provides fine-grained insight into where models break down across different dimensions.

Evaluating Depth of Understanding. We evaluate joint correctness across multiple cognitive tasks, penalizing partial understanding and preventing models from receiving credit for internally inconsistent reasoning. We define the *Social Understanding Score (SUS)*, which measures whether a model maintains and updates a correct social representation across a pivotal event while producing reasoning consistent with the observed change. We additionally define the *Social Agency Score (SAS)*, which measures whether a model can correctly anticipate how social interactions evolve over time, conditioned on grounded social understanding and reasoning. Let G denote the set of question groups, where each group $g \in G$ consists of a before-event comprehension question (if applicable), an after-event comprehension question, a reasoning question, and a prediction question. For each group, we define correctness indicators c_g , r_g , and p_g , where c_g indicates joint correctness of the before- and after-event comprehension questions, r_g indicates correctness of the reasoning question, and p_g indicates correctness of the prediction question.

$$\text{SUS} = \frac{1}{|G|} \sum_{g \in G} c_g \cdot r_g \quad \text{SAS} = \frac{1}{|G|} \sum_{g \in G} c_g \cdot r_g \cdot p_g \quad (1)$$

4 Dataset Constructions and Annotations

We develop an MLLM-assisted annotation and quality-control pipeline with 11 student annotators and 15 reviewers, producing 264 videos and 1,519 QA pairs (Figure 5).

4.1 Question Design

We identify pivotal moments in the video (plot twists) and design questions around them to test whether the model’s interpretation is correctly maintained up to that point and appropriately revised after a pivotal moment. Each question is timestamped, and during evaluation the video is trimmed to that timestamp, so the model only receives the context available up to that moment. This lets us directly compare the model’s interpretation with and without the pivotal event, isolating its effect on the model’s social understanding.



Figure 6: Input Video Context Distribution.

4.2 Quality Assurance

We implement the following to ensure the quality of our benchmark.

Human Baseline. To minimize bias from any individual annotator, we randomly sample 30 videos and have 15 reviewers (5 per age group, as shown in Figure 2) answer the corresponding questions. Their responses serve as the human baseline score against which all evaluated MLLMs are compared.

Temporal Dependency Filtering. To enforce long-horizon dependence, we verify that each question cannot be answered from a short local context. For every candidate QA pair, we extract a 20-second clip immediately preceding the queried timestamp and test whether the correct answer is recoverable using Gemini-3.1-Pro. Those that are recoverable are discarded, ensuring correctness requires reasoning across the full social interaction.

Visual Dependency Filtering. We pass the video transcriptions together with the QA pairs to ChatGPT-5.2 to identify and remove questions that can be answered correctly without visual information. This step ensures that correctness requires multimodal cues rather than explicit language.

Single Defensible Answer. Every question undergoes multiple rounds of cross-review by 3 annotators until unanimous agreement that there is a single defensible correct answer. Each question features several plausible answer options. However, we ensure that each question has only one defensible answer. For each question we also construct a variant in which the correct answer is replaced with "None of the answers are correct." This allows us to test whether models can not only identify the most plausible-sounding answer, but also correctly reject all distractors when none of them are valid. Each distractor must be unambiguously wrong, not just less plausible than the correct answer, so that "None of the above" is the only defensible choice in these variant items.

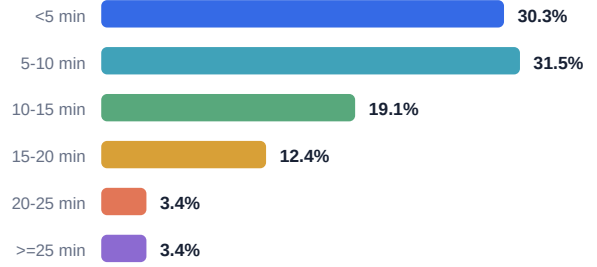


Figure 7: Input Video Length Distribution.

4.3 Video Collection

To prevent answers from being recovered via priors rather than video evidence, we exclude famous movies and instead collect publicly available videos containing plot twists across three domains, chosen for broad coverage of naturalistic human behavior: Scripted/Narrative Short Films (silent films, comedy sketches), Strategic/Competitive Interactions (sports clips, game shows, debates, competitions, social strategy games), and Conversational/Interpersonal Interactions (interviews, dating shows, everyday conversations).

5 Experiments

5.1 Experiment Setup

We evaluate a diverse suite of state-of-the-art MLLMs (see Table 2) on our 100 video test set, each in a zero-shot, single-turn setting with prompts kept as consistent as possible across models. All videos are standardized to 720p, and all models sampling rate is standardised to 1fps. To enable fair comparison regardless of audio support, each video is transcribed, and the transcript is passed as part of the prompt alongside corresponding timestamps. Each video is truncated at the question timestamp, requiring the model to reason over accumulated social context as it would unfold in a real interaction. To accommodate models' varying long-context capabilities, we use each model's recommended inference settings; malformed or unparsable outputs are treated as incorrect.

5.2 Main Results

Significant gap with human performance. Our results in Table 2 show that there still remains a significant gap between human and machine social intelligence. Even the best performing model, Gemini-3.1-Pro, reaches a social understanding score of 0.43 and a social agency score of 0.33 compared to the human accuracy of 0.85 and 0.78.

Table 2: Evaluation results on the *SMMU* test set (w/ subtitles). Best model results are **bolded**.

Model	SUS	SAS	Avg	Cognitive Tasks			Social Dimensions				
				ACS	ARS	APS	Pe	KS	Em	Re	In
Human Baseline	0.85	0.78	0.91	0.91	0.91	0.89	0.88	0.96	0.87	0.90	0.93
Gemini-3.1-Pro	0.43	0.33	0.75	0.73	0.78	0.78	0.62	0.94	0.72	0.86	0.79
Gemini-3.1-Flash-Lite	0.16	0.06	0.49	0.56	0.48	0.33	0.51	0.24	0.47	0.50	0.57
Kimi-K2.6	0.19	0.13	0.59	0.59	0.63	0.51	0.58	0.53	0.54	0.61	0.66
Seed-2.0-Pro	0.20	0.03	0.56	0.60	0.61	0.37	0.50	0.59	0.61	0.50	0.56
Qwen-3.7-Flash	0.05	0.00	0.40	0.44	0.42	0.25	0.32	0.41	0.39	0.50	0.40
Qwen-3.7-Plus	0.09	0.06	0.47	0.47	0.50	0.41	0.36	0.65	0.48	0.53	0.45
Qwen-3.5-Omni-Plus	0.18	0.07	0.51	0.56	0.56	0.28	0.51	0.47	0.51	0.47	0.53

SUS: Social Understanding Score (joint comprehension and reasoning); SAS: Social Agency Score (joint comprehension, reasoning, and prediction); Avg: Overall Average Accuracy; ACS: Average Comprehension Score; ARS: Average Reasoning Score; APS: Average Prediction Score; Pe: Perspective; KS: Knowledge State; Em: Emotions; Re: Relationships; In: Intent. Scores are micro-averaged over valid evaluated responses. Provider errors are excluded.

Conventional MCQ Benchmarks may overestimate models’ Social Intelligence. Traditional MCQ evaluation can overstate model performance because each question is scored independently. As shown in Table 2, average accuracy on independently scored questions across all models is substantially higher than joint metrics such as SUS and SAS. This large gap shows that strong performance on isolated questions does not necessarily reflect a coherent understanding across comprehension, reasoning, and prediction. For instance, a model may correctly recognize that a character is upset yet predict an action inconsistent with that emotional state, such as going out to celebrate. Similarly, it may correctly identify that one character is exploiting another while failing to recognize the evidence supporting that conclusion. This suggests that current models may select the most superficially plausible answer rather than grounding their interpretations in coherent, evidence-based reasoning. Our joint metrics, SUS and SAS, mitigate this limitation by evaluating whether models correctly update their comprehension as events unfold and whether their reasoning and predictions remain consistent with their evolving social understanding.

Correct comprehension does not guarantee correct reasoning or prediction. As shown in Table 3, models frequently fail to identify the evidence supporting a social interpretation or to anticipate what will happen next, even when their comprehension is correct. Correctly recognizing a social state therefore does not, by itself, demonstrate that a model has grounded its interpretation in the relevant audio, visual, and textual evidence or can reliably predict

Table 3: Conditional reasoning and prediction accuracy given correct or incorrect Comprehension.

Model	P(R C)	P(P C)	P(R ¬C)	P(P ¬C)
Gemini-3.1-Pro	0.82	0.93	0.75	0.55
Gemini-3.1-Flash-Lite	0.56	0.41	0.41	0.28
Kimi-K2.6	0.58	0.56	0.68	0.46
Seed-2.0-Pro	0.68	0.28	0.56	0.47
Qwen-3.7-Flash	0.33	0.15	0.44	0.29
Qwen-3.7-Plus	0.60	0.53	0.45	0.33
Qwen-3.5-Omni-Plus	0.64	0.42	0.48	0.13

C, R, and P denote End Comprehension, End Reasoning, and End Prediction, respectively. ¬C denotes an incorrect End Comprehension response.

how the interaction will unfold. Although most models perform better on downstream tasks following correct comprehension, this dependency is often weak. Gemini-3.1-Pro exhibits the strongest prediction dependency, with correct comprehension increasing prediction accuracy by 38.1 percentage points. In contrast, Kimi-K2.6, Seed-2.0-Pro, and Qwen-3.7-Flash perform similarly or better on at least one downstream task following incorrect comprehension. These results suggest that models’ representations of the social scene are only weakly coupled with how they answer different questions about the same video, consistent with reliance on post-hoc evidence matching or answer plausibility rather than coherent reasoning. Future work should evaluate and train for consistency across the complete reasoning chain and penalizing contradictions between its stated understanding, reasoning and subsequent inferences. We therefore argue that joint and conditional accuracy should be reported alongside marginal accuracy in social-intelligence benchmarks, as marginal scores alone cannot dis-



Figure 8: **Model Error Analysis.** Representative failure modes observed across models.

Table 4: Comprehension-transition outcomes and downstream failure rates across models. Only valid, error-free, non-stale responses are included. Correct Update is higher-is-better; all other metrics are lower-is-better. Values are reported as proportions.

Outcome	Gem	Gem-FL	Kimi	Seed	Qwen-F	Qwen-P	Qwen-O
Correct Update ↑	0.48	0.25	0.28	0.33	0.15	0.14	0.25
Premature Anticipation ↓	0.12	0.08	0.23	0.16	0.17	0.17	0.19
Failure to Update ↓	0.20	0.35	0.27	0.33	0.36	0.31	0.25
Reasoning Failure ↓	0.18	0.44	0.42	0.32	0.67	0.40	0.36
Prediction Inconsistency ↓	0.06	0.40	0.25	0.86	1.00	0.00	0.67

Gem: Gemini-3.1-Pro; Gem-FL: Gemini-3.1-Flash-Lite; Kimi: Kimi-K2.6; Seed: Seed-2.0-Pro; Qwen-F: Qwen-3.7-Flash; Qwen-P: Qwen-3.7-Plus; Qwen-O: Qwen-3.5-Omni-Plus. The transition metrics use Start-End pairs whose correct answers change. Reasoning Failure is measured conditional on correct End Comprehension. Prediction Inconsistency is measured conditional on successful social understanding. Denominators vary across models.

tinguish coherent compositional reasoning from independent question-level pattern matching.

5.3 Further Insights

Models are insufficiently sensitive to audiovisual evidence for social understanding. Most comprehension-transition errors arise from either *Premature Anticipation*, in which models select the correct answer before the supporting evidence (plot twist) appears, or *Failure to Update*, in which they retain an outdated interpretation after new ev-

idence (plot twist) is introduced. Although these failures occur in opposite directions, both indicate that model responses are weakly grounded in the visual, audio, and narrative evidence available at each stage of the video. Future work should therefore complement full-video evaluation with temporally truncated inputs that are progressively extended, enabling benchmarks to measure how models revise their understanding as new evidence becomes available rather than evaluating only their final interpretation after observing the entire video.

6 Conclusion

We introduced *SMMU*, a benchmark for evaluating long-horizon social understanding in MLLMs as videos unfolds over time. By jointly measuring comprehension updates, evidence-based reasoning, and prediction, *SMMU* reveals limitations that independently scored MCQs often obscure. Our results show that current frontier models remain substantially below human performance and fairly insensitive to pivotal audiovisual cues or produce conflicting answers. We hope *SMMU* supports the development of multimodal systems that reason more coherently, update their beliefs as new evidence emerges, and ground their social interpretations in the evolving audio-visual context.

Limitations

While *SMMU* provides a challenging benchmark for evaluating long-horizon multimodal social understanding, several limitations remain. First, although *SMMU* is designed to evaluate grounded social reasoning over extended audiovisual interactions, the benchmark remains limited in scale relative to large web-scale evaluation datasets. In particular, the benchmark contains 264 videos and 1519 question-answer pairs, and therefore may not fully capture the complete diversity of real-world social behavior and interaction patterns. Second, many social inferences are inherently subjective. Dimensions such as emotion, intent, perspective, and relationship dynamics may admit multiple plausible interpretations depending on cultural background, prior experiences, or social expectations. Consequently, benchmark annotations reflect annotator consensus rather than uniquely verifiable ground truth. Third, *SMMU* may encode cultural and linguistic biases. Social norms, sarcasm, politeness, embarrassment, hierarchy, and interpersonal signaling are often culturally dependent, and the benchmark primarily reflects the social conventions present in the collected video sources. Similarly, although transcript-only solvability is explicitly filtered, many questions still rely on English dialogue, prosody, and pragmatic cues, which may disadvantage models with weaker speech recognition or multilingual understanding capabilities. Finally, *SMMU* evaluates passive social understanding from prerecorded videos rather than interactive or embodied social intelligence. The benchmark does not evaluate real-time dialogue adaptation, multi-turn interaction, negotiation, collaborative planning, or long-term relationship modeling. We therefore view *SMMU* as a step toward evaluating grounded multimodal social reasoning rather than a complete measure of human-level social intelligence.

Ethical Considerations

SMMU is constructed from publicly available online videos used for academic research purposes. To respect platform access conditions and copyright restrictions, we do not redistribute raw video content directly. Instead, we release annotations, timestamps, and evaluation metadata. The benchmark is intended solely for non-commercial research use. Because the dataset contains real individuals appearing in publicly accessible videos,

it may contain identifiable faces, voices, and socially sensitive interactions. We therefore avoid collecting private or non-public content and focus on videos that are already broadly accessible online. Additionally, social interpretation is inherently subjective and may reflect annotator or cultural biases. To mitigate this, annotations were reviewed and refined through iterative quality control procedures.

Acknowledgments

This research is supported by cash and in-kind funding from NTU S-Lab and industry partner(s). This study is also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223- 0002). Additionally, we thank the following data annotators for their careful and diligent work on this project: Chuhan Mei, Ye Changyi, Laurel Leung, Ching Luk, Vivian Wen, Jialiang Song.

References

- Alibaba. 2026. [Qwen model family](#).
- Bytedance. 2026. [Seed model family](#).
- DeepMind. 2026. [Gemini model family](#).
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*.
- Google DeepMind. 2026. [Gemini 3.1 pro: A smarter model for complex tasks](#).
- Joy Paul Guilford. 1956. The structure of intellect. *Psychological bulletin*, 53(4):267.
- Francesca Happé, Jennifer L Cook, and Geoffrey Bird. 2017. The structure of social cognition: In (ter) dependence of sociocognitive processes. *Annual review of psychology*, 68:243–267.
- He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.
- Kimi. 2026. [Kimi model family](#).
- Dorit Kliemann and Ralph Adolphs. 2018. The social neuroscience of mentalizing: challenges and recommendations. *Current opinion in psychology*, 24:1–6.
- Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. Siv-bench: A video benchmark for social interaction understanding and reasoning. *arXiv preprint arXiv:2506.05425*.

- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.
- Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, and 1 others. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*.
- Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. 2024. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*.
- Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24879–24902.
- Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.
- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3762–3780.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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.
- Bytedance Seed. 2026. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.
- Robert L Thorndike and Saul Stein. 1937. An evaluation of the attempts to measure social intelligence. *Psychological bulletin*, 34(5):275.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, and 1 others. 2025b. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*.
- Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2025. VideoChat-R1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*.
- Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2025. Mm-instructeval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. *Information Fusion*, 122:103204.
- Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.
- 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*.

Appendix

A Candidate Answer Prompt

The following prompt is used for candidate answer generation in the multiple-choice evaluation setting. The model is instructed to analyze the video and answer a question.

B Question-Answer Refinement Prompt

The following prompt is used to refine the wording of the questions.

C Annotation Design

Answer Choice Construction. Each question in SMMU is accompanied by multiple plausible answer choices designed to evaluate fine-grained social understanding rather than simple elimination strategies. Distractors are constructed to remain highly plausible within the broader video context and, in many cases, factually valid in isolation, while still being definitively incorrect under the specific social interaction or queried context. Rather than enforcing a fixed number of distractors per question, annotations prioritize quality and plausibility, with each question containing at least one correct answer and multiple socially plausible alternatives. To assist the annotation process, annotators were permitted to use MLLM-generated suggestions as inspiration for distractor creation; however, all distractors were manually reviewed and refined, as automatically generated candidates were frequently overly obvious or insufficiently grounded in the social context.

Cognitive Dependency Structure. Questions in SMMU are organized into progressive cognitive triplets consisting of Comprehension, Reasoning, and Prediction tasks. Reasoning questions are only included when a corresponding Comprehension question is present, enforcing a dependency between lower-level understanding and higher-level inference. Prediction questions are included selectively, only when subsequent social behaviour or future outcomes are sufficiently meaningful and inferable from the preceding interaction context.

Temporal Grounding. To ensure temporally grounded social reasoning and prevent hindsight bias, all annotations are constructed exclusively using information available up to the queried timestamp. During evaluation, videos are truncated precisely at the selected timestamp such that models

A. Candidate Answer Prompt

You are helping an annotator answer a multiple-choice question about a video clip.

Use only evidence that is visible or audible in the clip.

Select exactly one option label from the list below.

Return only valid JSON. Do not include markdown, explanation, reasoning text, or any extra keys.

Use this exact shape: `{"selected_label": "A"}`.

Question:

{question}

Return only the letter corresponding to the best answer in the required schema.

only observe video context preceding or coinciding with the queried moment. Consequently, questions are designed such that the relevant social state remains implicit at the timestamp while still being inferable from subtle multimodal cues already present in the interaction. Later dialogue or events that explicitly reveal the answer are not permitted to serve as the primary justification for the annotation. For example, if a character later explicitly states an emotion or intention, the earlier question must still remain answerable from preceding behavioural and contextual evidence alone.

Character Identification. SMMU additionally maintains a structured character library for identifying recurring individuals within each video. When character names are explicitly mentioned in the video, annotations use those names directly. Otherwise, characters are identified using visually distinctive attributes such as clothing, occupation, or physical appearance (e.g., “the man in the blue hoodie” or “the police officer standing beside the car”). Character references are required to remain sufficiently specific to avoid ambiguity among multiple individuals appearing within the same scene.

D Dataset Usage and Licensing

SMMU is constructed from publicly available online videos used for academic research purposes. To respect platform access conditions and copyright restrictions, we do not redistribute raw video

B. Question-Answer Refinement Prompt

You are an expert at rewriting multiple-choice questions for clarity and evaluation quality.

You are given a Multiple Choice Question (MCQ) with answer choices. Your task is to rewrite the question and choices while preserving their original meaning.

Requirements:

1. Improve clarity, grammar, and fluency. Correct any typos or awkward phrasing.
2. Do NOT change the semantic meaning of the question or any answer choice.
3. Ensure the question is precise, unambiguous, and self-contained.
4. Keep all answer choices grammatically parallel and similar in length and style.
5. Avoid introducing new information or removing necessary context.

content directly. Instead, we release annotations, timestamps, and evaluation metadata. The benchmark is intended solely for non-commercial research use.

Plot Synopsis: A young girl at risk of being trafficked finds an unlikely ally in a peeping tom, the receptionist, who helps her escape using a ladder.



The Receptionist



Intent

[16:21.07] Which of the following correctly describes what the receptionist's intent is at this point in the video?

- A. He intends to create a discreet escape route for the woman.
- B. He intends to reach the window so that he can remove the hidden camera.
- C. He intends to help the woman's next customer enter the room without passing through reception.
- D. He intends to repair the window after noticing that it cannot be secured from inside.
- E. He intends to retrieve an object that the woman has placed outside the window.
- F. He intends to stage a break-in so that the disappearance of the woman will not be connected to him.
- G. He intends to gain access to the room because the woman has locked its door from inside.
- H. He intends to enter the room to take advantage of the woman.

[16:21.07] Which of the following best explains why the receptionist intends to create a discreet escape route for the woman, or how the receptionist achieves it?

- A. He feels remorseful for the woman who is being abused.
- B. After arguing with the handler about unpaid hotel charges, he helps the woman escape primarily to retaliate against him.
- C. After discovering that the woman has seen the hidden camera, he helps her leave in exchange for her promise not to expose him.
- D. After receiving a note in which the woman asks for help escaping, the receptionist feels compelled to assist her.

Plot Synopsis: An impoverished child grows up to become the doctor who saves the life of the man who helped him 30 years ago.



The Doctor



Relationship

[02:18.33] Which of the following correctly describes the relationship between the doctor and the food stall owner at this point in the video?

- A. The owner helped the doctor as an impoverished child.
- B. The food stall owner is the doctor's father.
- C. The doctor has no previous personal connection with the owner but regards treating him without charge as a way of continuing the owner's pattern of generosity.
- D. The owner previously supported the doctor throughout his childhood and helped pay for the education that allowed him to enter medicine.
- E. The doctor knows the owner as a generous community member and waives the bill after learning that the owner has spent many years feeding people without charge.
- F. The doctor is the impoverished child's mother
- G. The doctor is the impoverished child's father

[02:18.33] Which of the following best explains what indicates that the owner helped the doctor as an impoverished child?

- A. The bill says that the treatment was paid 30 years ago.
- B. The doctor takes a personal interest in the owner's condition and continues overseeing his treatment after the operation.
- C. The thirty-year time jump allows the boy from the opening scene to have reached approximately the doctor's age.
- D. The doctor eliminates the medical expenses after the daughter advertises the urgent sale of her father's business.
- E. The daughter finds an old photograph showing her father standing beside the doctor when the doctor was still a child.
- F. The boy's experience of being unable to obtain medicine for his mother may have inspired him to pursue a career in healthcare.
- G. The doctor recognizes the owner's name in the hospital records and tells a nurse that the man regularly fed him when he was homeless.

Plot Synopsis: An elderly man carries a bed board into an elevator and accidentally gets stuck, unable to press the elevator buttons.



Emotions

[01:49.00] Which of the following correctly describes what the old man is feeling at this point in the video?

- A. Stressed
- B. Calm
- C. Joyous
- D. Indifferent
- E. Confused
- F. Frustrated

[01:49.00] Which of the following best explains what causes the old man to feel the following emotion: Stressed?

- A. The man is stuck in the elevator.
- B. The premise of the question is wrong. The man is calm and waiting for the lift to go down.
- C. The man cannot fit the huge gate into the elevator.
- D. The elevator broke down.

[01:49.00] Which of the following is most likely to occur next?

- A. Call for help using phone
- B. Press the button to reopen the doors of the elevator
- C. Press the button to get to the floor he needs to go to.
- D. Press a random floor to exit the elevator
- E. Press the emergency button in the elevator next to all the buttons

Figure 9: SMMU Additional Question Samples. SMMU structures evaluation as a chain of cognitive tasks: Comprehension (left), Reasoning (middle), and Prediction (right).