SOCIAL GENOME: Grounded Social Reasoning Abilities of Multimodal Models

Leena Mathur[*†1], **Marian Qian**[*1], **Paul Pu Liang**[2], **Louis-Philippe Morency**[1]
Carnegie Mellon University[1], Massachusetts Institute of Technology[2]
{lmathur,marianq,morency}@cs.cmu.edu, ppliang@mit.edu

Abstract

Social reasoning abilities are crucial for AI systems to effectively interpret and respond to multimodal human communication and interaction within social contexts. We introduce SOCIAL GENOME, the first benchmark for fine-grained, grounded social reasoning abilities of multimodal models. SOCIAL GENOME contains 272 videos of interactions and 1,486 human-annotated reasoning traces related to inferences about these interactions. These traces contain 5,777 reasoning steps that reference evidence from visual cues, verbal cues, vocal cues, and external knowledge (contextual knowledge external to videos). SOCIAL GENOME is also the first modeling challenge to study external knowledge in social reasoning. SOCIAL GENOME computes metrics to holistically evaluate semantic and structural qualities of model-generated social reasoning traces. We demonstrate the utility of SOCIAL GENOME through experiments with state-of-the-art models, identifying performance gaps and opportunities for future research to improve the grounded social reasoning abilities of multimodal models.

1 Introduction

Humans rely on *social reasoning* to interpret and navigate everyday interactions (Gagnon-St-Pierre et al., 2021). This form of reasoning is a core competency of social intelligence (Kihlstrom and Cantor, 2000; Conzelmann et al., 2013), occurs with specialized neural and cognitive systems (Cao et al., 2024; Read et al., 2013), and involves integrating information over time from multimodal behaviors such as gestures, language, and prosody (Morency, 2010; Read and Miller, 2014; Liang et al., 2024). Multimodal cues are often *fine-grained* (e.g., a fleeting glance), *interleaved* (e.g., a shrug followed by a sigh), and *context-dependent*, requiring *external knowledge* of contextual information to be interpreted accurately (Hechter and Opp, 2001).

[*]equal contribution, †corresponding author

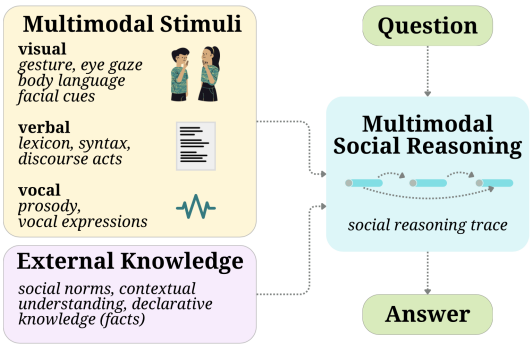

Figure 1: Reasoning over multimodal social interactions involves extracting, integrating, and referencing evidence from multiple behavioral modalities, as well as information from external knowledge.

Developing algorithms for multimodal social reasoning will be essential to advance artificial intelligence (AI) systems with social intelligence (Mathur et al., 2024). When AI systems reason about human social interactions, it is important for systems to have the ability to generate explanations with accurate, *grounded* references to fine-grained multimodal behaviors and external knowledge concepts informing inferences. Figure 1 visualizes these aspects of multimodal social reasoning. This capability is especially important for AI systems reasoning about interactions in high-stakes domains, such as healthcare and assistive robots.

Progress towards improving multimodal social reasoning in models has been limited by a lack of evaluation tasks – measuring a capability is an essential first step towards advancing it. To address this challenge, we introduce **SOCIAL GENOME**, the first benchmark for grounded multimodal social reasoning that includes 272 videos of face-to-face interactions and 1,486 human-annotated reasoning traces explaining inferences about social information in these videos. Across these traces, SOCIAL GENOME contains 5,777 social reasoning steps. Each reasoning step is tagged with the modality of

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24880–24903
November 4-9, 2025

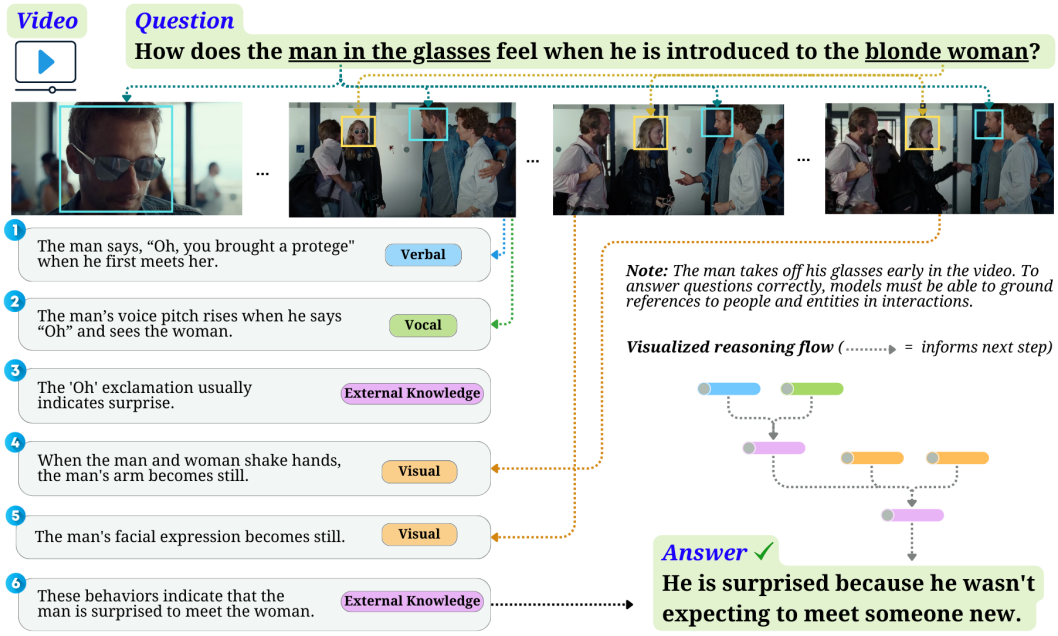

Figure 2: A sample reasoning trace from the SOCIAL GENOME benchmark. Reasoning traces in SOCIAL GENOME contain fine-grained, multimodal social cues and references to external knowledge informing the social inference. Social reasoning traces produced by humans can contain complex reasoning paths (sample visualized above) that reference and build upon multimodal evidence and external knowledge across temporal segments of interactions.

information being referenced: *visual*, *verbal*, and *vocal* cues from social interactions in videos, and *external knowledge* of contextual information that human annotators used to perform social inferences (information external to stimuli in videos). Reasoning traces in SOCIAL GENOME are *dense* with references to over 11,000 entities (people, objects, concepts), over 5000 multimodal cues, and over 2,900 external knowledge observations. SOCIAL GENOME is the first social reasoning benchmark[1] that includes external knowledge and *dense* reasoning traces. A sample human-annotated reasoning trace is visualized in Figure 2.

This paper defines metrics to holistically assess semantic and structural aspects of model-generated social reasoning traces. We demonstrate the utility of SOCIAL GENOME by using these metrics to distill insights regarding social reasoning capabilities and limitations in state-of-the-art (SOTA) models. For example, we find that models struggle to perform well under both zero-shot and in-context learning (ICL) settings, demonstrating the significant challenge of building this understudied form of reasoning in models. Our findings contribute novel insights regarding gaps and opportunities for future research to improve grounded social reasoning abilities of multimodal models.

[1] cmu-multicomp-lab.github.io/social-genome

2 Background

Prior research on social reasoning in models has primarily focused on the ability of models to interpret text-based social scenarios and perform question-answering (QA) tasks about characters' motivations, intents, and actions; Social IQa remains a key unimodal benchmark in this area (Sap et al., 2019). SOTA language models can accurately perform a majority of the inferences in Social IQa, but a gap remains between model and human performance (Sap et al., 2022; Shapira et al., 2024). SOTA models have also struggled with text-based QA tasks that probe competencies relevant to social reasoning, specifically theory-of-mind to interpret the goals and beliefs of characters (Le et al., 2019; Shapira et al., 2024; Ullman, 2023; Kim et al., 2023). Crowd-sourced knowledge bases of norms (Forbes et al., 2020; Ziems et al., 2023) have been useful to inform social reasoning research.

The ability of models to reason about multimodal social interactions, in particular *face–to–face*, *embodied*, *real-world* social interactions, has been comparatively understudied. Key benchmarks include the video QA tasks of Social-IQ 1.0 (Zadeh et al., 2019) and Social-IQ 2.0 (Wilf et al., 2023); both examine model QA accuracy when answering questions about social interactions in videos. SOTA models have struggled to perform well on

Benchmark	Social Reasoning Task Focus	Reasoning Traces	Fine-Grained Evaluation	External Knowledge
SOCIAL GENOME (this paper)	✓	✓	✓	✓
SOCIAL-IQ 1.0 (Zadeh et al., 2019)	✓	✗	✗	✗
SOCIAL-IQ 2.0 (Wilf et al., 2023)	✓	✗	✗	✗
TVQA (Lei et al., 2018)	✗	✗	✗	✗
MOVIEQA (Tapaswi et al., 2016)	✗	✗	✗	✗
MEMOR (Shen et al., 2020)	✗	✗	✗	✗
EGO4D-SOCIAL (Grauman et al., 2022)	✗†	✗	✗	✗

Table 1: SOCIAL GENOME enables the study of fine-grained, grounded social reasoning in multimodal models. ✓ = provided; ✗ = not provided. †Ego4D has a social signal perception task, which differs from a social reasoning task.

Social-IQ 2.0 (Xie and Park, 2023; Pirhadi et al., 2023; Li et al., 2024c; Agrawal et al., 2024; Chen et al., 2024). Prior focus on QA *accuracy* to assess social reasoning ability has not enabled researchers to study the extent to which models can effectively reference *fine-grained* multimodal cues and *external knowledge* informing inferences. Models with high accuracy on QA tasks can perform poorly at generating valid or comprehensive reasoning traces (Jhamtani and Clark, 2020; Gu et al., 2023), motivating the creation of benchmarks with fine-grained evaluation that goes beyond QA tasks. We introduce SOCIAL GENOME as the first benchmark to study grounded, fine-grained social reasoning in multimodal models. Table 1 summarizes the novelty of SOCIAL GENOME relative to prior human-centered video understanding tasks.

3 Building SOCIAL GENOME

3.1 Sourcing Seed Videos and Questions

SOCIAL GENOME contains 272 seed videos and 1486 questions adapted from the SOCIAL-IQ 2.0 dataset (Wilf et al., 2023) (details in Appendix A.1). Videos include real-world face-to-face interactions (1 min per video, ∼4.5 hours); questions probe behaviors, emotions, and cognitive states of individuals and groups. SOCIAL GENOME introduces a new set of 1486 human reasoning traces with 5700+ steps that answer these questions.

3.2 Task Notation

Given a video V, a question Q about social interactions in the video, and answer options $A = \{A_{\text{correct}}, A_{\text{incorrect}_1}, A_{\text{incorrect}_2}, A_{\text{incorrect}_3}\}$, a model performing the SOCIAL GENOME task must generate a reasoning trace $R = \{e_1, e_2, \ldots, e_n\}$, where each reasoning step e_i represents a single piece of evidence contributing toward the social inference to select an answer A_{a} from A. Each reasoning step e_i must be tagged with two attributes: (1) a modality tag $m_i \in \{\textit{visual, verbal, vocal, } n/a\}$ indicating the communication modality of the evidence and (2) an external knowledge tag $k_i \in \{\textit{yes}, \textit{no}\}$, indicating whether the evidence references external knowledge of contextual information. This task to generate R and answer question Q evaluates a model's ability to extract and reference multimodal aspects of human communication and knowledge informing social inferences. Given the input tuple (V, Q, A), each model performing the SOCIAL GENOME task will produce an output tuple (A_{a}, R). Metrics in SOCIAL GENOME study the social inference accuracy of A_a and the semantic and structural aspects of social reasoning in R.

3.3 Social Reasoning Trace Annotations

Human Annotation Given a video V, question Q, and answer options A, annotators read Q and A, watched V, and wrote reasoning trace R. Annotations were collected with an IRB-approved Prolific study (details in Appendix A.3).

Grounded and Fine-Grained Behaviors Humans build upon low-level observations of fine-grained behaviors (e.g, shifts in body language) and high-level, top-down processing (e.g., implicit situational knowledge) when interpreting social scenes (Baird and Baldwin, 2001; Bodenhausen and Morales, 2013). Annotators were instructed to reference any *low-level* and *high-level* evidence they relied upon to answer questions: for example, low-level evidence might be "the woman takes a step back with her mouth wide open (visual cue)" and high-level evidence that interprets that low-level cue might be "the woman is surprised (external knowledge regarding how 'surprise' might manifest"). For each step, annotators tagged the modality referenced (visual, verbal, vocal, n/a).

Grounded External Knowledge Annotators tagged each reasoning step with yes or no to indicate whether external knowledge was referenced.

External knowledge includes contextual norms, cultural expectations, and prior understanding of social commonsense (Forguson and Gopnik, 1988) that goes beyond stimuli in the video. For example, if a man raises his arm and the annotator recognizes his movement as a "high five," the identification of the gesture is based on external knowledge of social gestures. If a reasoning step focuses solely on noting a low-level behavioral signal (e.g., "the man raised his voice rapidly"), this evidence would not be viewed as containing external knowledge.

Ensuring Annotation Quality Trained experts validated each Prolific annotation. They watched each video, read each QA tuple, and read the annotation to ensure that traces represented valid reasoning, had correct modality and external knowledge tags, and referred to relevant information. Cases of incomplete annotation or deviation from instructions were fixed (details in Appendix A.4).

3.4 Dataset Statistics

SOCIAL GENOME contains 1486 human-annotated reasoning traces with 5,777 total steps, 3.89 $\pm$ 1.68 steps per trace (min of 1 step and max of 10 steps), 43 $\pm$ 26 words per trace, and 11$\pm$ 5 words per step. Reasoning steps draw on multimodal evidence; 44% of steps reference visual cues, 27% reference verbal cues, and 17% reference vocal cues. Overall, 77% of traces reference at least one visual cue, 63% reference at least one verbal cue, and 47% reference at least one vocal cue.

External knowledge plays a critical role in SOCIAL GENOME: 51% of reasoning steps referenced external knowledge, with each trace referencing an average of 2 pieces of external knowledge. With spaCy named entity recognition (NER) (Honnibal et al., 2020), we found 11,253 entities (people, objects, concepts) mentioned, with 7.6 unique entities and 2.23 emotions referenced per reasoning trace, demonstrating the high *density* of annotations. Additional details regarding dataset statistics are in Section 7 and Appendix A.2.

3.5 Social Reasoning Metrics and Statistics

We develop metrics to evaluate *semantic* and *structural* aspects of social reasoning traces generated by models performing tasks in the SOCIAL GENOME benchmark. Collectively, these metrics reveal strengths and weaknesses in model social reasoning and multimodal grounding abilities and the extent to which model traces differ from human reasoning. This multi-dimensional evaluation mitigates against models hacking individual metrics to achieve higher scores. For each sample, we compute the following metrics between model reasoning trace R_M with n steps $e_1...e_n$ and the corresponding human trace R_H with m steps $h_1...h_m$.

Accuracy Measures accuracy of the model-generated answer by comparing it to ground truth. Human annotator accuracy on SOCIAL GENOME is 0.85 (Appendix A.5). Higher values indicate stronger social inference ability (max value is 1).

Similarity-Trace (S_{trace}) Measures the high-level semantic similarity between R_M and R_H:

$$S_{\text{trace}} = \frac{\langle \mathbf{R_M}, \mathbf{R_H} \rangle}{\|\mathbf{R_M}\| \cdot \|\mathbf{R_H}\|}$$

where $\mathbf{R_M}$ is the aggregate embedding of evidence steps in R_M and $\mathbf{R_H}$ is the aggregate embedding of evidence steps in R_H. The embedding model `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019), selected for its efficiency and accuracy, was used to embed evidence steps for this and other semantic similarity metrics. Higher values indicate stronger alignment in semantic information between R_M and R_H (max value is 1).

Similarity-Step (S_{step}) Measures the fine-grained semantic similarity between R_M and R_H. For each step e_i in R_M, the metric identifies its closest semantic step h_j in R_H. The final metric is the mean of these maximum similarity values:

$$S_{\text{step}} = \frac{1}{n} \sum_{i=1}^{n} \max_j \frac{\langle \mathbf{e_i}, \mathbf{h_j} \rangle}{\|\mathbf{e_i}\| \cdot \|\mathbf{h_j}\|}$$

where $\mathbf{e_i}$ is the embedding of evidence step e_i and $\mathbf{h_j}$ is the embedding of evidence step h_j. Higher values reflect stronger alignment in semantic information between fine-grained steps of evidence in R_M and R_H (max value is 1).

Similarity-Num Steps (S_{num}) Measures the number of steps in R_M with a similarity above threshold τ, when compared to any step in R_H:

$$S_{\text{num}} = \sum_{i=1}^{n} \mathbb{1} \left(\max_j \frac{\langle \mathbf{e_i}, \mathbf{h_j} \rangle}{\|\mathbf{e_i}\| \cdot \|\mathbf{h_j}\|} > \tau \right)$$

where $\mathbb{1}(\cdot)$ is the indicator function and $\tau = 0.6$ (empirically selected). Higher values indicate more semantically-aligned evidence between R_M and R_H (max value is n).

DifferenceSequence (DS) Measures structural similarity between R_M and R_H using the respective modality sequences S_M and S_H from the reasoning traces (e.g., ["visual", "external knowledge"]). A similarity score based on edit distance, adapted from the Levenshtein distance, is computed between S_M and S_H (Appendix B.2). Higher values of DS indicate greater structural similarity between R_M and R_H (max value is 1).

EmotionMetric Measures the alignment of emotional content in R_M and R_H. This metric extracts sets of emotions referenced by R_M and R_H (instructions in Appendix B.3) and computes the overlap in sets. Higher values can indicate stronger emotional alignment between R_M and R_H (max value is emotion set size in R_H).

All Modality Steps Measures the overlapping number of unique modalities in R_M and R_H. Higher values indicate that R_M had more overlap with modalities referenced by R_H (max value is the number of unique modalities in R_H).

Visual Steps Measures the number of steps in both R_M and R_H with *visual* evidence, to evaluate how closely R_M aligns with R_H (max value is number of visual steps in R_H).

Verbal Steps Measures the number of steps in both R_M and R_H with *verbal* evidence, to evaluate how closely R_M aligns with R_H (max value is the number of verbal steps in R_H).

Vocal Steps Measures the number of steps in both R_M and R_H with *vocal* evidence to evaluate how closely R_M aligns with R_H (max value is the number of vocal steps in R_H).

External Knowledge Steps Measures the number of steps in both R_M and R_H with *external knowledge* evidence, to evaluate how closely R_M aligns with R_H (max value is the number of external knowledge steps in R_H).

NumSteps (NS) Measures the absolute difference in the number of reasoning steps between R_M and R_H. Lower values indicate stronger alignment in length between model and human chains (value of 0 indicates that R_H and R_H are the same length).

3.6 SOCIAL GENOME ICL Training Set

To create samples for ICL experiments, we randomly sampled 16 questions from unique videos in the *training* set of SOCIAL-IQ 2.0 and collected reasoning trace annotations in the same format as annotations in Section 3.3. ICL experiments in Section 4 are conducted by providing each model with different numbers of training samples $k \in \{0, 1, 2, 4, 8, 16\}$, before the model is given an input tuple (V, Q, A) and generates a sequence of tokens with an answer A_a and reasoning trace R.

4 Social Reasoning Experiments

We use SOCIAL GENOME to study the performance of multimodal video understanding models in fine-grained, grounded social reasoning. These models have exhibited SOTA performance on video understanding tasks and take videos as input. We tested 2 closed-source models and 5 open-source models: Gemini-1.5-Flash (Team et al., 2024), GPT-4o (OpenAI), LLaVA-Video and LLaVA-Video-Only (Zhang et al., 2024c), LongVA (Zhang et al., 2024a), Video-ChatGPT (Maaz et al., 2023), and VideoChat2 (Li et al., 2023a). Models have different architectures, pretraining data, and fine-tuning tasks, and models generated reasoning traces and answers for all samples (details in Appendix C). LLaVA-Video-Only answered questions, but did not generate valid traces that could be studied; this model does not appear in trace-related metrics.

4.1 Quantitative Results and Insights

Figure 3 visualizes each model's average performance for our 12 metrics[2], with human baselines in gray. Results tables (Tables 4, 5, 6) are in Appendix C. Key findings are discussed below.

Social Inference Accuracy *Human social inference ability is substantially higher than all models*, as seen in results from the **Accuracy** metric. Gemini-1.5-Flash and GPT-4o achieve the highest accuracies of 74.4% and 71.0% respectively ($k = 0$), approximately 10-15% lower than human annotator accuracy (85.3%), answering ∼30% of questions incorrectly. *Closed-source models outperform open-source models in social inference*. The highest-performing open-source model was LLaVA-Video at 62.9% ($k = 0$). Gemini-1.5-Flash and GPT-4o are much larger than open-source models, suggesting that increased model *scale* is useful, but not sufficient, for social inference. Video-

[2]Visualized metrics are normalized to [0,1] by human baselines, except for those with an upper bound of 1 by definition (Accuracy, Similarity-Trace, Similarity-Step, DifferenceSequence) and absolute measures (NumSteps).

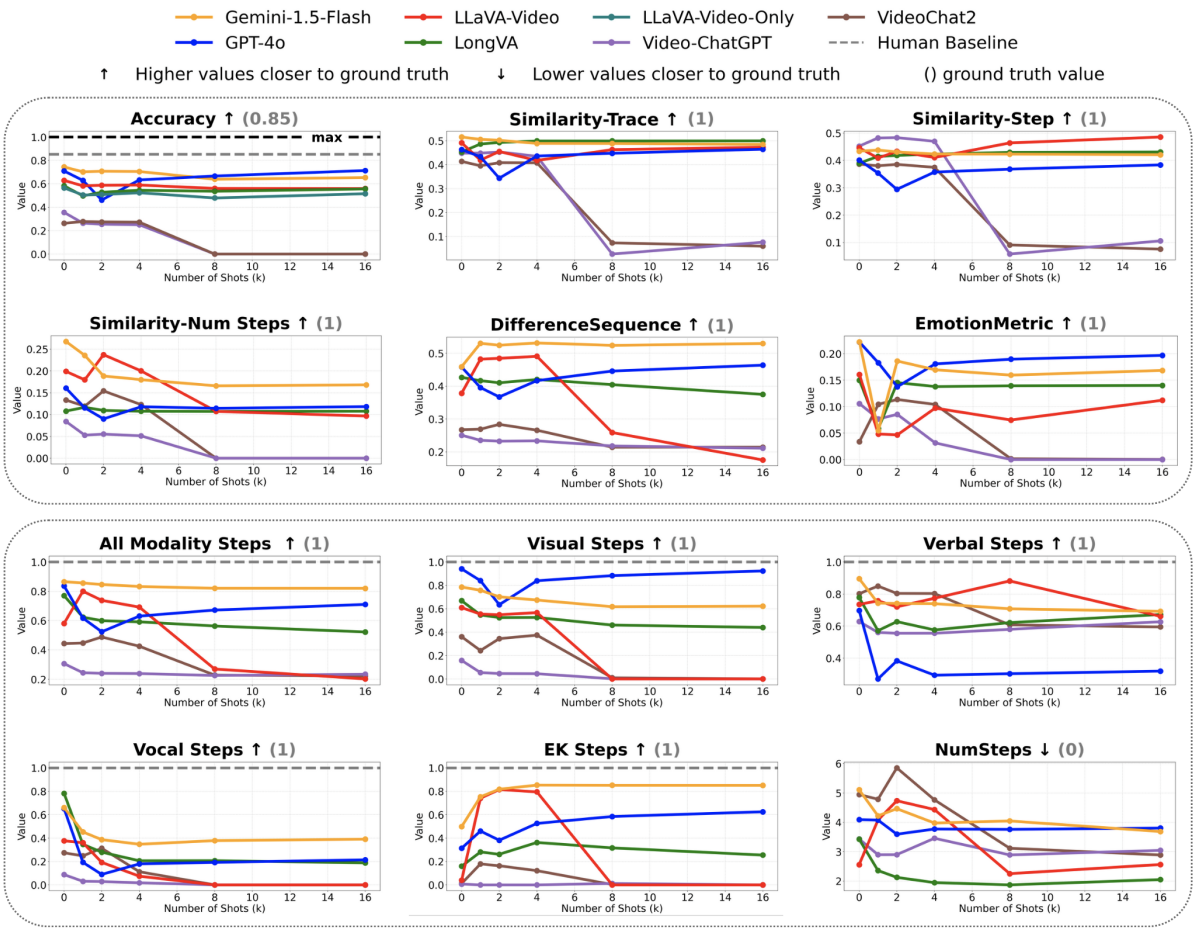

Figure 3: Performance of models across number of few-shot ICL samples (k) and ground truth noted in gray. The first six metrics focus on social inference accuracy, semantic similarity, and structural similarity between model and human reasoning traces. The final six metrics focus on fine-grained multimodal evidence and external knowledge referenced by models, in comparison to evidence referenced by humans. *With these metrics,* SOCIAL GENOME *enables a holistic study of multimodal, grounded social reasoning.*

ChatGPT and VideoChat2 perform 30-40% lower than other open-source models across all values of k. These models have the smallest context length, constraints which may influence performance.

Social inference accuracy for models decreased as the number of few-shot ICL samples increased, with the exception of GPT-4o, which demonstrated a slight improvement at $k = 16$. Few-shot ICL conditions language models on tasks by providing examples of inputs and outputs (Liu et al., 2024) and can be viewed as a form of *inductive reasoning*, as the model is tasked with inferring generalizable rules from a set of examples. This technique has improved model reasoning abilities in domains such as mathematics and code generation (Dong et al., 2024; Zhou et al., 2022; Patel et al., 2023), which have *explicit* rules and formal structure (Galotti, 1989). In contrast to these domains, social reasoning often operates with *implicit* rules, less formal structure (Perkins, 1989) and ambiguity in premises (Mathur et al., 2024). Our findings suggest that few-shot ICL may not be an effective approach to elicit multimodal social reasoning abilities. Additional experiments using SOCIAL GENOME samples as a form of supervision for models (discussed in Appendix D) demonstrate that chain-of-thought prompting did not improve model accuracy, and models struggled to perform inferences relying on implicit and contextual knowledge. Our finding that few-shot ICL is insufficient to elicit multimodal social reasoning aligns with findings from unimodal experiments on SOCIAL-IQA and TOMI datasets (Le et al., 2019; Sap et al., 2019, 2022; Kim et al., 2023).

Semantic Alignment *It is challenging for models to generate social reasoning traces with high semantic alignment to human reasoning*, as seen in the results from the **Similarity-Trace**, **Similarity-**

Step, and **Similarity-Num Steps** metrics. For Similarity-Trace, only Gemini-1.5-Flash achieved slightly above 50% (k = 0), and LongVA outperformed Gemini-1.5-Flash after k = 4. For Similarity-Step, no models achieved above 50%, but Video-ChatGPT outperformed all models for $k \in \{0, 1, 2, 4\}$ (5-12% higher than GPT-4o and 2-5% higher than Gemini-1.5-Flash). For Similarity-Num Steps, Gemini-1.5-Flash achieved the highest performance (27%), far below ground truth. Low performance can be explained by failure to reference cues that humans reference ("Fine-Grained Grounding" metrics below). Performance on these metrics did not improve as k increased.

Structural Alignment *Model reasoning traces tend to reference multimodal evidence in different amounts and orders than humans*, as seen in the results from the **DifferenceSequence** metric. Model-generated sequences of multimodal evidence that were most structurally-aligned with human sequences were from Gemini-1.5-Flash (0.53 at k = 4) and LLaVA-Video (0.49 at k = 4), both far below the maximum alignment value (1). As k increased, DifferenceSequence metric scores increased for Gemini-1.5-Flash, LLaVA-Video (up to k = 4), and GPT-4o (after k = 2). These findings suggest that structured samples of human social reasoning, as introduced by SOCIAL GENOME, can be useful when conditioning models to generate reasoning traces with more human-like structure.

Emotion Alignment *It is challenging for models to generate social reasoning chains with high emotional alignment with human reasoning*, as seen in the results from the **EmotionMetric**. All models achieved less than 22%, far below ground truth. While the highest scores were achieved by Gemini-1.5-Flash and GPT-4o at k = 0, scores from several models steadily improved after k = 2.

Fine-Grained Multimodal Grounding The final 6 metrics study evidence referenced by model reasoning traces. The **AllModality Steps** metric serves as a proxy for how well models refer to fine-grained, multimodal cues and external knowledge. As seen in AllModality Steps results, *all models referenced fewer pieces of multimodal evidence and external knowledge than human reasoning.*

While we discuss modality-specific findings below, our use of SOCIAL GENOME to study SOTA models is currently limited by these models' modality processing abilities. The Gemini-1.5-Flash API reports that audio is processed on the backend, but other models do not process audio. However, despite the absence of audio, models referenced verbal and vocal evidence *inferred* from visual frames. For example, LongVA generated vocal evidence about a woman's "tone suggesting frustration," after generating visual evidence about the woman's face appearing dissatisfied. As new models are developed in the coming years with abilities to jointly process video and audio, SOCIAL GENOME will continue to be applicable to study model abilities in grounded multimodal social reasoning.

Visual Steps: *Closed-source models exhibited a strong ability to reference visual evidence*, with GPT-4o and Gemini-1.5-Flash closer to ground truth than open-source models. Performance on this metric for all models was highest at k = 0.

Verbal Steps: *The ability of models to reference verbal evidence showed substantial variation*. Gemini-1.5-Flash exhibited the strongest ability to reference verbal cues (0.90 at $k = 0$). Model performance on this metric did not improve as k increased, and GPT-4o referenced substantially fewer verbal cues in comparison to other models.

Vocal Steps: *The ability of models to reference vocal evidence was substantially lower than human ability*. LongVA referenced more vocal evidence than other models at k = 0. Model performance on this metric did not improve as k increased.

External Knowledge Steps: *The ability of models to reference external knowledge in social reasoning traces was lower than humans*. In contrast to trends observed in other modality step metrics, we find that providing additional few-shot samples improved the ability of Gemini-1.5-Flash, GPT-4o, LongVA (up to k = 4), and LLaVA-Video (up to k = 4) to reference external knowledge. These findings suggest that human social reasoning traces can be used to condition models to ground social reasoning with external knowledge references.

Num Steps: *Model reasoning traces varied in length and contained more steps than human traces*. LLaVA-Video and LongVA generated model reasoning traces that were most aligned with the length of human traces. Providing additional few-shot samples improved the ability of models to align with human social reasoning trace length.

4.2 Human and Qualitative Evaluation

We conducted human evaluation of model reasoning traces. Trained annotators analyzed 48 samples from all models, with model names and k values

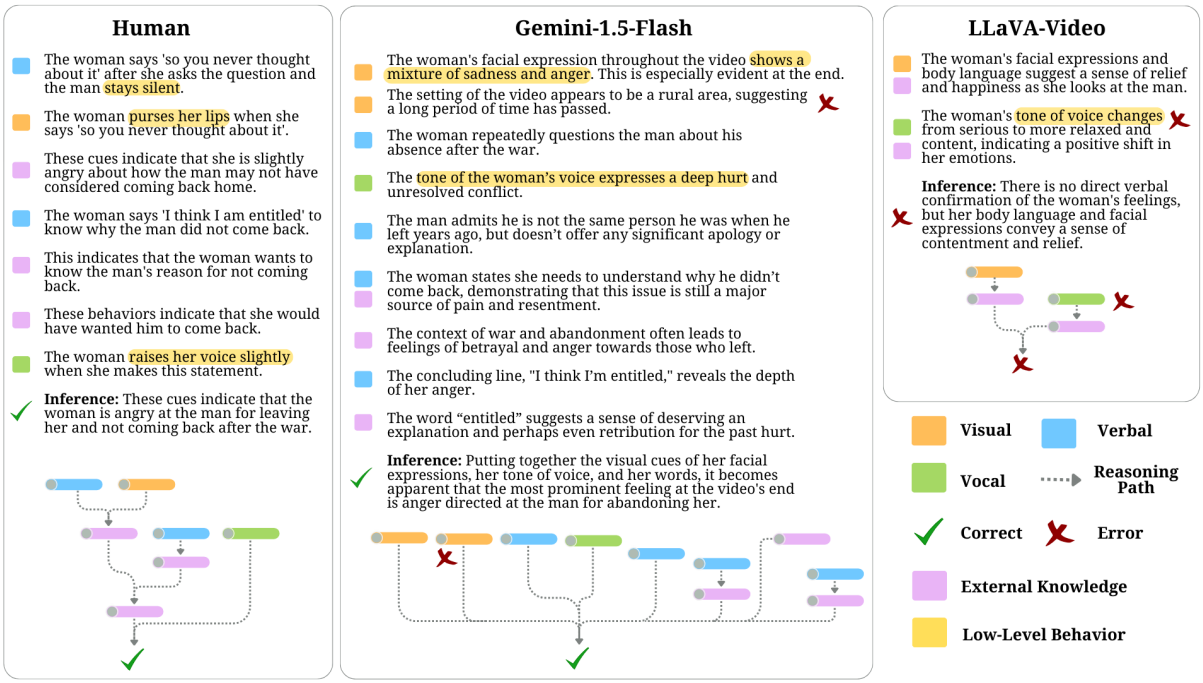

Figure 4: Representative social reasoning traces from a Human Annotator, Gemini-1.5-Flash, and LLaVA-Video. These examples illustrate the grounded social reasoning abilities and reasoning structures across humans and models.

anonymized. Traces were rated to assess references to low-level cues (*fine-grained*), information cross-referenced across steps (*compositional*), relevant evidence (*comprehensive*), *correctness* of modality tags, and *validity* of reasoning (details in Appendix E). Annotator agreement was computed with Cohen's Kappa κ (Cohen, 1960).

Reasoning traces from Gemini-1.5-Flash and GPT-4o received the highest ratings for *fine-grained* ($\kappa = 0.87$), *compositional* ($\kappa = 0.92$), *comprehensive* ($\kappa = 0.94$), and *valid* ($\kappa = 0.94$) reasoning, followed by LLaVA-Video and LongVA, with VideoChatGPT and VideoChat2 rated lowest. These findings validate model performance trends in Section 4.1. Modality tag correctness across samples was 98%, and metrics correlated with human judgements ($R^2 > 0.75$ for Gemini-1.5-Flash semantic metrics, details in Appendix E), supporting the validity of benchmark processing.

Evidence and Error Propagation Figure 4 illustrates the strong ability of Gemini-1.5-Flash to reference and integrate multimodal cues and external knowledge. In contrast, LLaVA-Video did not reference low-level cues and based its reasoning upon an incorrect premise which led to an incorrect inference. The human trace referred to *fine-grained behaviors* (e.g., lip movements) that are *not* present in the model traces, yet do influence the scene interpretation. Metrics in Section 3.5 serve as proxies to estimate these types of information gaps between human and model reasoning. Our findings motivate future work on model training and architectures that better capture fine-grained cues and handle error propagation in reasoning.

Hierarchical Social Reasoning Figure 4 shows the *hierarchical structure* of human social reasoning traces, in which low-level cues (e.g., brief lip movement) are referenced, combined, and re-interpreted as *intermediate evidence* for further reasoning. "Forking" reasoning structures are common for humans interpreting everyday situations, unlike the linear "long chain" structures for formal reasoning in domains such as mathematics (Perkins, 1989; Galotti, 1989). Compared to human traces, model traces show flatter structures which have the potential to overlook intermediate evidence. Our findings motivate future work to train models capable of more hierarchical social reasoning.

5 Conclusion

We introduce SOCIAL GENOME, the first benchmark for fine-grained, grounded social reasoning abilities of multimodal models. Reasoning traces contributed by SOCIAL GENOME include multimodal cues and external knowledge concepts that humans find useful when performing social infer-

ences. We define metrics to assess semantic and structural aspects of reasoning traces and contribute novel insights regarding gaps and opportunities to improve the grounded social reasoning capabilities of multimodal models. Future AI systems reasoning about social interactions must be able to *ground* reasoning in concrete multimodal evidence and external knowledge concepts. SOCIAL GENOME serves as a first step towards studying and advancing this form of reasoning in AI systems.

6 Limitations

Social Reasoning in Natural Language The current scope of SOCIAL GENOME focuses on studying model-generated social reasoning traces in *natural language*. This scope is necessary and relevant to contexts that require AI systems to generate natural language explanations of social inferences – for example, a healthcare agent or hospital robot reasoning about human nonverbal behaviors during a nurse-patient social interaction. However, an open question remains regarding the extent to which natural language can effectively represent the nuances of human social interactions and social reasoning (Mathur et al., 2024). It is possible that both humans and models verbalizing social reasoning through natural language are not fully capturing *why* they came to certain inferences (Turpin et al., 2023). Several lines of work in reasoning have operated in the *latent space* of models instead of natural language (Hao et al., 2024; Geiping et al., 2025), and we believe SOCIAL GENOME informs and motivates future work to develop techniques to study social reasoning in the latent space.

Video Lengths The videos in SOCIAL GENOME each have a length of ∼1 minute, consistent with the lengths of existing video understanding benchmarks (e.g., SOCIAL-IQ 1.0 and SOCIAL-IQ 2.0 have 1-minute samples (Wilf et al., 2023; Zadeh et al., 2019), TVQA has 1.3 minute samples (Lei et al., 2018), and MEmoR has 30 second samples (Shen et al., 2020)). *Social reasoning regularly occurs in micro-social and shorter-term contexts*; humans make split-second inferences about emotions (Nook et al., 2015), social behaviors and gestures (Beattie and Aboudan, 1994), and personality (Lin et al., 2021), among other social phenomena. The interactions in SOCIAL GENOME videos contain rich, nuanced social signals and multimodal behavioral dynamics that require social reasoning to interpret. Our paper demonstrates that current state-of-the-art models struggle to interpret 1-minute social interactions. The length of our videos is not a technical limitation of our research. We would like to motivate the need for community-driven curation of longer-form datasets in future years.

Scope of the Study Videos in SOCIAL GENOME have interactions in English, and annotators were required to be proficient in English. The study was scoped within these constraints, consistent with prior multimodal video understanding tasks (Table 1). A multilingual and multicultural data collection was not within the scope of this research. Our paper motivates future research in multimodal social reasoning that includes a community-driven curation of interaction data across sociocultural contexts.

7 Ethics

Ethical Annotation We curated annotations from videos in existing publicly available datasets. We hired workers from Prolific to annotate reasoning traces. All workers received fair compensation for their annotation ($12 per hour, pro-rated). Worker privacy and confidentiality were respected, with no identifiable information stored. Further details on Prolific annotation are in Appendix A.3.

Bias Considerations Annotators in the original SOCIAL-IQ 2.0 dataset, from which we sourced seed videos, used terms such as "man" in alignment with annotator perception of gender. In SOCIAL GENOME we did not frame judgements about gender identity of individuals based on these annotations. In SOCIAL GENOME, 45% of samples refer to women, and 17% make no reference to gender. Samples involving women do not have reasoning traces that refer more frequently to emotion words (r = 0.033). We find no substantial difference in model performance across gender; for example, Gemini-1.5-Flash social inference accuracy is 74.9% for samples solely involving women, 73.4% for sampling solely involving men, 73.6% for samples referring to multiple genders, and 77% for samples that do not specify gender.

Environmental Statement Experiments used a single A100 GPU, a carbon footprint of 1.24 kgCO2e, and an energy consumption of 3.72 kWh[3].

Risks for Social Reasoning in AI Social reasoning abilities are essential for future AI systems to

[3]http://calculator.green-algorithms.org

effectively work *with* and *alongside* humans. SOCIAL GENOME has the potential to support the research community in studying and advancing these capabilities in AI systems. We envision AI systems using social reasoning to enhance human autonomy, health, and well-being. However, these technologies exist with potential risks in amplifying toxicity (Zhou et al., 2023), surveillance, and manipulation. We support broader research and policy efforts to mitigate against misuse and potential harms of socially-intelligent AI.

Acknowledgments

Leena Mathur is supported by the NSF Graduate Research Fellowship Program under Grant No. DGE2140739. This material is based upon work partially supported by National Institutes of Health awards R01MH125740, R01MH132225, and R21MH130767. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred. Figure 2 includes icon material available from https://icons8.com.

References

Aviral Agrawal, Carlos Mateo Samudio Lezcano, Iqui Balam Heredia-Marin, and Prabhdeep Singh Sethi. 2024. Listen then see: Video alignment with speaker attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2018–2027.

Jodie A Baird and Dare A Baldwin. 2001. Making sense of human behavior: Action parsing and intentional inference.

Geoffrey Beattie and Rima Aboudan. 1994. Gestures, pauses and speech: An experimental investigation of the effects of changing social context on their precise temporal relationships. *Semiotica*, 99(3-4):239–272.

Galen V Bodenhausen and Javier R Morales. 2013. Social cognition and perception. *Handbook of psychology*, 5:225–246.

Runnan Cao, Julien Dubois, Adam N Mamelak, Ralph Adolphs, Shuo Wang, and Ueli Rutishauser. 2024. Domain-specific representation of social inference by neurons in the human amygdala and hippocampus. *Science Advances*, 10(49):eado6166.

Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2024. Through the theory of mind's eye: Reading minds with multimodal video large language models. *arXiv preprint arXiv:2406.13763*.

Georgios Chochlakis, Niyantha Maruthu Pandiyan, Kristina Lerman, and Shrikanth Narayanan. 2024. Larger language models don't care how you think: Why chain-of-thought prompting fails in subjective tasks. *arXiv preprint arXiv:2409.06173*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Kristin Conzelmann, Susanne Weis, and Heinz-Martin Süß. 2013. New findings about social intelligence. *Journal of Individual Differences*.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670. Association for Computational Linguistics.

Lynd Forguson and Alison Gopnik. 1988. The ontogeny of common sense. *Developing theories of mind*, pages 226–243.

Émilie Gagnon-St-Pierre, Marina M Doucerain, and Henry Markovits. 2021. Reasoning strategies explain individual differences in social reasoning. *Journal of Experimental Psychology: General*, 150(2):340.

Kathleen M Galotti. 1989. Approaches to studying formal and everyday reasoning. *Psychological bulletin*, 105(3):331.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Yuling Gu, Oyvind Tafjord, and Peter Clark. 2023. Digital socrates: Evaluating llms through explanation critiques. *arXiv preprint arXiv:2311.09613*.

Xiao-Yu Guo, Yuan-Fang Li, and Gholamreza Haffari. 2023. Desiq: Towards an unbiased, challenging benchmark for social intelligence understanding. *arXiv preprint arXiv:2310.18359*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Michael Hechter and Karl-Dieter Opp. 2001. Social norms.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. *arXiv preprint arXiv:2010.03274*.

John F Kihlstrom and Nancy Cantor. 2000. Social intelligence. *Handbook of intelligence*, 2:359–379.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

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

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023a. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

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

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023b. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960.

Yunxin Li, Xinyu Chen, Baotain Hu, and Min Zhang. 2024c. Llms meet long video: Advancing long video question answering with an interactive visual adapter in llms. *arXiv preprint arXiv:2402.13546*.

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

Chujun Lin, Ü Keleş, and Ralph Adolphs. 2021. Four dimensions characterize comprehensive trait judgments of faces. *Nature Communications*, 12(1):5168.

Emmy Liu, Graham Neubig, and Jacob Andreas. 2024. An incomplete loop: Deductive, inductive, and abductive learning in large language models. *arXiv preprint arXiv:2404.03028*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in AI agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, Miami, Florida, USA. Association for Computational Linguistics.

Leena Mathur, Maja Mataric, and Louis-Philippe Morency. 2023. Expanding the role of affective phenomena in multimodal interaction research. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 253–260.

Louis-Philippe Morency. 2010. Modeling human communication dynamics [social sciences]. *IEEE Signal Processing Magazine*, 27(5):112–116.

Erik C Nook, Kristen A Lindquist, and Jamil Zaki. 2015. A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, 15(5):569.

OpenAI. Gpt-4o system card.

Arkil Patel, Siva Reddy, Dzmitry Bahdanau, and Pradeep Dasigi. 2023. Evaluating in-context learning of libraries for code generation. *arXiv preprint arXiv:2311.09635*.

David N Perkins. 1989. Reasoning as it is and could be: An empirical perspective.

Mohammad Javad Pirhadi, Motahhare Mirzaei, and Sauleh Eetemadi. 2023. Just ask plus: Using transcripts for videoqa. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3082–3085.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Stephen J Read and Lynn C Miller. 2014. On the dynamic construction of meaning: An interactive activation and competition model of social perception. In *Connectionist models of social reasoning and social behavior*, pages 27–68. Psychology Press.

Stephen J Read, Chadwick J Snow, and Dan Simon. 2013. Constraint satisfaction processes in social reasoning. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 964–969. Psychology Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 493–502.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

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

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.

Baijun Xie and Chung Hyuk Park. 2023. Multi-modal correlated network with emotional reasoning knowledge for social intelligence question-answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3075–3081.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of

large language models: A survey. *arXiv preprint arXiv:2406.04244*.

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.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

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

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

A SOCIAL GENOME Dataset Appendix

A.1 SOCIAL GENOME Data Sourcing

All videos in SOCIAL GENOME were sourced from publicly-available SOCIAL-IQ 2.0 modeling challenge (Wilf et al., 2023). We obtained permission from the authors to access the original *test set* videos and question-answer tuples from this dataset (275 videos and 1514 QA Tuples). Our use of these videos aligns with Social-IQ 2.0 repository's MIT license and intended research purpose. We chose to use these test set videos as candidate seed videos to build SOCIAL GENOME because the answers to questions posed about these videos *have not been released online*, reducing chances of benchmark contamination (Jacovi et al., 2023; Xu et al., 2024; Deng et al., 2024). Videos, questions, and answer options are available and open to the community, consistent with the permissive MIT License for SOCIAL-IQ 2.0. Scripts to access our SOCIAL GENOME dataset and prepare model outputs into submissions for our leaderboard are available under an MIT license. Our leaderboard, accessible from the project website[4], supports researchers in computing metrics on SOCIAL GENOME.

We note that prior papers that test models on SOCIAL-IQ 2.0 have used the validation set, not the test set (Xie and Park, 2023; Pirhadi et al., 2023; Guo et al., 2023; Li et al., 2024c; Agrawal et al., 2024; Chen et al., 2024). Therefore, we do not directly compare prior works' validation set performance with our results on the test set.

During manual inspection of each video and QA tuple, we filtered out QA tuples with answer options containing ambiguity – if at least 2 annotators judged a question to have ambiguous answer options (at least two plausibly-correct answers), we discarded the question. The resulting SOCIAL GENOME set contained 272 videos and 1486 QA tuples. These samples contain face-to-face dyadic and multi-party social interactions and questions that probe understanding of affective states, causal social dynamics, and social events.

A.2 SOCIAL GENOME Entities

Across the 1486 samples and 11,253 entities (people, objects, concepts) mentioned in SOCIAL GENOME, we visualize the distribution of non-human entities (objects, concepts) in Figure 5, with an average of 7.6 unique entities referenced per human-annotated reasoning trace. Figure 6 visualizes entities most frequently mentioned by human annotators (not including words such as "man" and "woman" repeated from question statements). As seen in Figure 6, human annotators constructing social reasoning traces focused on multimodal aspects of social interactions, with evidence spanning *vocal* (e.g., "tone", "voice"), *visual* (e.g., "eyes", "head", "body"), and *verbal* (e.g, "words", "conversation") cues. These observations support the perspective that interpreting and reasoning about real-world social interactions requires an integration of multimodal information.

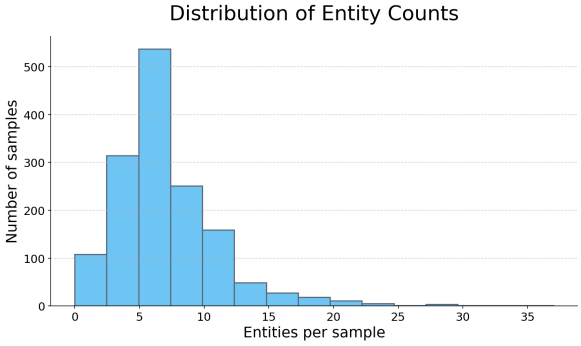

Figure 5: Distribution of entity counts in SOCIAL GENOME samples.

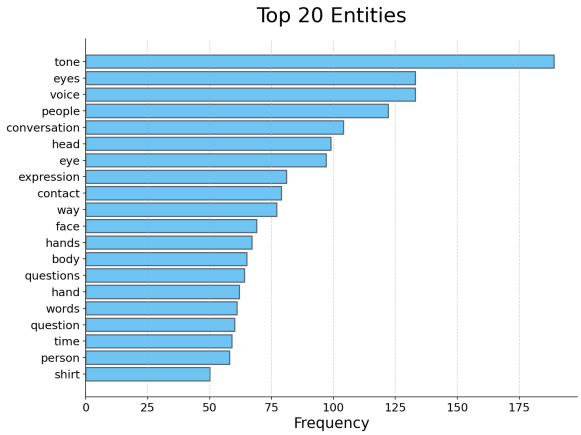

Figure 6: Top entity counts mentioned by human annotators in SOCIAL GENOME reasoning traces.

Annotators mentioned an average of 2.23 emotions per trace, indicating that affective content informed their reasoning when interpreting interactions (Mathur et al., 2023). Emotional constructs most frequently mentioned by annotators were the following: happy, serious, surprised, excited, nervous, calm, confused, angry, annoyed, comfortable, sarcastic, and positive.

[4] cmu-multicomp-lab.github.io/social-genome

A.3 SOCIAL GENOME Annotation

Annotators were recruited on the Prolific[5] platform to perform an IRB-approved study with informed consent regarding our data use. Screens for annotators were employed to ensure that participants were adults, had access to a desktop to watch videos and perform annotations, were fluent in English, were based in the United States, completed at least a high school education, and had high prior task approval rates on Prolific (97-100% completion). Each annotator watched 2 videos and provided reasoning traces for all questions associated with each video (approximately 10 questions per annotation). Instructions for annotators are listed in Figure 7. Annotators were compensated at $12 per hour prorated. This process was time-intensive, with each annotation taking approximately ∼25 minutes.

Before running the Prolific study, we ran 4 iterations of annotation instructions with pilot groups to refine instructions. We experimented with instructions that had annotators provide reasoning traces without knowing the correct answer and while knowing the correct answer. We found that providing annotators with the correct answer option did not change the detail, structure, or coherence of the reasoning traces generated. Therefore, we chose to provide annotators with the correct answer option (Figure 7) and focus the large-scale Prolific data collection on obtaining detailed reasoning traces, instead of additional QA answer responses.

A.4 SOCIAL GENOME Validation

After obtaining human-annotated reasoning chains from Prolific, we conduct validation to ensure the validity and correctness of annotations. At least 2 annotators evaluated each trace, and complex cases were evaluated by at most 3 annotators to watch each video, read each question and set of answer options, and check whether chains (1) represented valid reasoning paths, (2) had correct evidence tags for *visual, verbal, vocal* and *external knowledge*, and (3) comprehensively referred to relevant low-level social information in videos. If an annotator did not complete an annotation in a satisfactory manner (e.g., incomplete reasoning chain, minimal effort), the task was returned to them on Prolific. If annotations could be rapidly fixed (e.g., changing an incorrect modality tag from "visual" to "vocal"), the authors performed this fix themselves without sending the annotation task to the annotator.

[5]https://app.prolific.com

A.5 SOCIAL GENOME Human Accuracy

For each video in SOCIAL GENOME, two human annotators on Prolific watched the video, read each question, and selected one of four provided answer options. Annotators were paid $12 per hour (pro-rated) for completing each study. For each question, annotators also had the option to indicate whether or not they were “uncertain" about the answer option they selected, and the authors examined these samples to ensure all QA tuples in SOCIAL GENOME had correct answers, to avoid the situation in which a sample could have more than one plausibly-correct answer option. Human annotators on Prolific achieved an accuracy of 85.3% on the 1486 questions in SOCIAL GENOME. Inter-annotator agreement among Prolific annotators was positive (Cohen's κ = 0.60) (Cohen, 1960), and answer correctness was confirmed independently by authors, as described earlier.

B SOCIAL GENOME Metrics Appendix

B.1 Embeddings for Semantic Similarity

The metrics Similarity-Trace, Similarity-Step, and Similarity-Num Steps are computed with embeddings from the `all-MiniLM-L6-v2` from Sentence-BERT (Reimers and Gurevych, 2019). We found this embedding strong, efficient, and useful for our task; as future embedding models are enhanced and released in the coming years, the SOCIAL GENOME framework allows for the embedding model called by semantic similarity metrics to be updated.

B.2 DifferenceSequence Metric

For model reasoning chain R_M with modality sequence S_M and human reasoning chain R_H with modality sequence S_H, the DifferenceSequence (DS) metric is computed as a normalized similarity score by adapting the Levenshtein distance (Levenshtein, 1966) between S_M and S_H:

$$DS = 1 - \frac{\text{Levenshtein}(S_M, S_H)}{|S_M| + |S_H|}$$

$\text{Levenshtein}(S_M, S_H)$ is the following:

$$= \begin{cases} |S_M|, \text{if } |S_H| = 0 \\ |S_H|, \text{if } |S_M| = 0 \\ \min \Big[\text{Levenshtein}(S_M[:-1], S_H[:-1]) + \delta, \\ \quad \text{Levenshtein}(S_M[:-1], S_H) + 1, \\ \quad \text{Levenshtein}(S_M, S_H[:-1]) + 1\Big] \text{otherwise} \end{cases}$$

Annotation Instructions

In this study, you will be watching short videos that contain social interactions. You will be briefly explaining your thought process along several dimensions while answering questions about each video. Your answers will be entered through a Google spreadsheet. Participants in this study must have the ability to watch videos and write logical text responses in English describing their step-by-step reasoning while interpreting videos.

If you give your consent to participate in this study, please select "I agree": `<I agree>`

(If the annotator declines, they move to a page that states: "As you do not wish to participate in this study, please return your submission on Prolific by selecting the 'Stop without completing' button.")

In the provided spreadsheet (link below), you will see 2 videos linked in the "Video URL" column and a set of corresponding questions and answers for each video. The yellow-highlighted box indicates the correct answer for each question. Your goal is to do the following:

1. Click the blue URL next to each set of questions in the "Video URL" column. Watch the video and pay attention to the behaviors of people interacting.
2. For each question associated with each video, provide the reasoning chain in order to answer the question:
   (a) Read the question, read the answer options, and note the correct answer (highlighted in yellow).
   (b) In the "Evidence" column, describe the reasoning steps you took to reach the correct answer. Put each step in a separate row in the spreadsheet. (If applicable) steps should start with low-level observations (behavioral information from the video OR external knowledge) and build up to higher-level concepts.
3. For each piece of evidence, in the same row, fill out the "Modality" column drop-down menu to indicate whether your evidence is from visual information (facial movements, gestures, anything vision-related, etc), vocal information (e.g., audio/speech), verbal information (e.g., the actual content of words being spoken), or external knowledge (e.g., cultural norms, your own understanding of what a behavior means).
4. Please aim for 5-10 reasoning steps per question (if applicable).

We will look at the spreadsheet during our manual review of submissions. All questions must have evidence presented with the answer selected: `<link to the spreadsheet>`

Figure 7: Sample instructions provided to annotators on Prolific to provide SOCIAL GENOME reasoning traces. Annotators entered information into a structured spreadsheet with links to videos, questions, answer options, and columns to enter reasoning traces, modality tags, and external knowledge tags.

with $\delta = \mathbb{1}(S_M[-1] \neq S_H[-1])$, where $\mathbb{1}(\cdot)$ returns 1 if the final elements differ and 0 otherwise. We use an implementation[6] that treats a substitution as equivalent to one insertion plus one deletion, making the distance effectively an InDel distance. We compute the minimum number of edits that are needed to transform one sequence to another. Higher edit distances indicate that more edits are needed to align sequences (more dissimilarity).

Therefore, the overall DS similarity metric between S_M and S_H can range from 0 (maximum number of edits required) to 1 (minimum number of edits required). Higher DS values indicate greater structural similarity between the sequences S_M and S_H, with respect to the type and order of modality evidence being referenced.

B.3 Emotion Named Entity Recognition

We perform NER with spaCy (Honnibal et al., 2020). In spaCy's NER v3 configuration, we broadly defined an `emotion` label as a "`description of how a person is feeling`". To avoid identifying words like "feels" as entities, we also passed an example into the spaCy NER configuration that explicitly labeled "feels" as not an emotion entity (e.g., "`She feels sad because her friend didn't come with her`"). There are an average of 2.23±1.63 emotion entities referenced per human reasoning chain.

B.4 Chain Processing for Metrics

Computing metrics requires a standardized format for model-generated reasoning chains. Several model generations (in particular, the generations from open-source models) required processing before metrics could be computed.

We first parsed generations from models by splitting each generation based on its structure, such as the presence of line breaks, numbering, or sentences. For example, if a model generated a multi-sentence response, but did not include line breaks

[6] https://rapidfuzz.github.io/Levenshtein/index.html

Model Prompt Template

Trace: Having a model provide a *reasoning trace* in the response (remove the "few-shot examples" for zero-shot experiments at $k = 0$):

> Watch the video and answer questions about social information in the video, following the format of the examples below.
>
> <FEW-SHOT EXAMPLE with Question and Answer Options>
>
> Which of these is the correct answer? Output either A, B, C, or D. Provide the reasoning steps you took to get to this answer. Give 5 to 10 reasoning steps in bullet point form. Tag each bullet point step as visual, vocal, and verbal modality from the frames or external knowledge. Use the format: The correct answer <insert answer>.The correct answer is <FEW-SHOT RESPONSE WITH ANSWER AND REASONING TRACE>.
>
> Which of these is the correct answer? Output either A, B, C, or D. <CURRENT QUESTION AND ANSWER OPTIONS> Provide the reasoning steps you took to get to this answer. Give 5 to 10 reasoning steps in bullet point form. Tag each bullet point step as visual, vocal, and verbal modality from the video or external knowledge. Use the format: The correct answer is <insert answer>.

No Trace: Having a model answer the question with *no* reasoning trace (remove the "few-shot examples" for zero-shot experiments at $k = 0$):

> Watch the video and answer questions about social information in the video, following the format of the examples below.
>
> <FEW-SHOT EXAMPLE with Question and Answer Options>
>
> Which of these is the correct answer? Output either A, B, C, or D. Use the format: The correct answer <insert answer>.The correct answer is <FEW-SHOT CORRECT ANSWER, FEW-SHOT REASONING CHAIN>.
>
> Which of these is the correct answer? Output either A, B, C, or D. <CURRENT QUESTION AND ANSWER OPTIONS> Use the format: The correct answer is <insert answer>.

Note: For GPT-4o prompts, we found that we needed to replace the word *video* with *frames* to avoid error messages associated with video processing (e.g., "*I am unable to view or analyze video frames directly. However, I can help answer questions based on descriptions or provide general information. Let me know how I can assist you!*"). Other models did not have this issue.

Figure 8: Information on model prompts to obtain reasoning traces and inferences from models.

or numbering within the response, we would split this model output by individual sentences (e.g., splitting on the "." character).

We, then, parsed through each step and remove any steps that simply repeated the question or answer choices. We also removed phrases such as "reasoning step", "reasoning", or "the correct answer", as those phrases were often in steps like "The correct answer is A." or "Below are the reasoning steps:". In addition, during in-context learning experiments, we found that some model generations (in particular, GPT) would contain repetitions of sample chains within the model's response, leading the response to contain 2-3 reasoning chains. We automatically processed these responses by only taking the final reasoning chain out of these multiple chains and checking that this final chain was answering the original question.

Models were tasked with tagging modalities for each step of their generated reasoning chains. To validate this process, we automatically checked whether each step included *visual, vocal, verbal* or *external knowledge* tags. However, models sometimes failed to tag modalities for their chains. To automatically handle these cases, we employed `GPT-4o-mini` (Hurst et al., 2024) to tag modalities. We note that the models that needed this additional validation step were VideoChat2, VideoChatGPT, and LLaVA-Video; generations from these models did not have any modality tagging for the majority of their chains. The authors manually inspected a subset of GPT-generated modality tags for model-generated reasoning traces to verify accuracy.

C Social Genome Model Appendix

C.1 Model Information

Experiments were conducted with multimodal models that were selected for their SOTA performance on various video understanding tasks and have the ability to take a full video as input (2 closed-source models and 5 open-source models): Gemini 1.5 Flash (Team et al., 2024), GPT-4o (OpenAI), VideoChat2 (Li et al., 2023a), Video-ChatGPT (Maaz et al., 2023), LLaVA-Video (Zhang et al., 2024c), LLaVA-Video-Only (Zhang et al., 2024c), LongVA (Zhang et al., 2024a). We summarize the models below, and Appendix Table 2 lists characteristics of these models: context length, tokens per frame, training max frame, parameter count, and backbone. Figure 8 describes the prompts given to models. Our benchmark allows experiments with any model that processes multimodal language and video input and outputs text, allowing Social Genome to be used as a benchmark over time to study social reasoning. Experiments were conducted with one A100 GPU.

VideoChat2 The VideoChat2 model (Li et al., 2024b) has an architecture with UMT-L vision encoder (Li et al., 2023b), QFormer and Vicuna-7B v0 language model, has 7B parameters, can process up to 16 frames, and was trained with instruction tuning on a collection of 34 tasks spanning conversations, captions, visual question-answering, reasoning, and classification.

Video-ChatGPT The VideoChat-GPT model (Maaz et al., 2023) has an architecture built on top of LLaVA, with a CLIP vision encoder (Radford et al., 2021) and a Vicuna-7B v1.1 language model. The VideoChat-GPT model can process up to 100 frames, and was trained on video instruction pairs from the VideoInstruct100K dataset.

LLaVA-Video The LLaVA-Video model `LLaVA-Video-7B-Qwen2` (Zhang et al., 2024c) has an architecture with a SigLIP SO400M vision transformer and Qwen2 language model, has 7B parameters, can process up to 110 frames, and was trained on mixture of single image, multi-image, and video tasks from the LLaVA-Video-178K and LLaVA-OneVision datasets (Li et al., 2024a).

LLaVA-Video-Only The LLaVA-Video-Only model `LLaVA-Video-7B-Qwen2-Video-Only` is identical to the LLaVA-Video model, with the exception of the training data (Zhang et al., 2024c). LLaVA-Video-Only was solely trained on the LLaVA-Video-178K dataset.

LongVA The LongVA model `LongVA-7B-DPO` (Zhang et al., 2024a) aligns a unified multimodal transformer (UMT) with QFormer and aligns this visual encoder with a Qwen2 7B language model. LongVA was trained on visual instruction-following datasets and multimodal document data and has a context length of over 200,000 visual tokens; this longer context length was achieved by extending the context length of the language backbone to train on long text samples, before performing multimodal alignment and additional training to transfer this ability to the multimodal domain.

GPT-4o The GPT-4o model is a closed-source model from OpenAI (Hurst et al., 2024). The tech-

Model Property	VideoChat2	Video-ChatGPT	LLaVA-Video	LLaVA-Video-Only	LongVA	GPT-4o	Gemini 1.5
Context	2K	2K	32K	32K	224K	128K	1M
Max Frames	16	100	110	110	2000	-	-
Parameters	7B	7B	7B	7B	7B	-	-
Backbone	Vicuna-v0	Vicuna-1.1	Qwen2	Qwen2	Qwen2-Extended	-	-

Table 2: Information about models tested on SOCIAL GENOME: VideoChat (Li et al., 2023a), VideoChat-GPT (Maaz et al., 2023), LLaVA-Video (Zhang et al., 2024b), LLaVA-Video-Only (Zhang et al., 2024b), LongVA (Zhang et al., 2024a), GPT-4o (OpenAI), Gemini 1.5 Flash (Team et al., 2024). All information is completed based on current public reports and repositories.

nical report for GPT-4o refers to this model as "omnimodal" with the ability to accept inputs from text, audio, image, and video and generate outputs with text, audio, and image. The API access to this model supports video frame inputs and text inputs.

Gemini-1.5-Flash The Gemini-1.5-Flash model is a closed-source model from Google. The API access to this model supports video inputs and text inputs, up to a context length of approximately 1 million tokens. Gemini-1.5-Flash was distilled from the larger Gemini-1.5-Pro sparse mixture-of-experts transformer (Team et al., 2024).

C.2 Model Generation Notes

We note observations here on model generations. VideoChat generations for $k \in 0, 1, 2, 4$ produced full sentences explaining reasoning traces, however the generation quality eroded for $k \in 8, 16$. Several samples from these settings of k were repeated words and short phrases (e.g.," `the the the...and and and and`").

Similarly, VideoChat-GPT generations for $k \in 0, 1, 2, 4$ produced full sentences explaining reasoning, however the generation quality eroded for $k \in 8, 16$. Samples from these settings of k were repeated short words and letters (e.g.,"`or or or, or`" and "`B ( ( ( ( ( B`").

LLaVA-Video generations for $k \in 0, 1, 2, 4, 8, 16$ produced full sentences explaining reasoning, however the generations as k increased in $k \in 4, 8, 16$ began to answer fewer questions. LLaVA-Video-Only answered questions, but did not generate reasoning traces; this model was discussed in Section 4 solely for the social inference accuracy metric.

These model generation challenges were not observed for LongVA or Gemini-1.5. GPT-4o initially generated "`I'm sorry, I can't assist with that.`" as one of the reasoning trace steps for several questions, before answering the question.

D Auxiliary Experiments

D.1 Social Inference Accuracy and Reasoning Trace Lengths

We hypothesized that models may perform worse on inferences that primarily rely on *implicit* cues and *contextual* knowledge. One proxy for this reliance is the length of a human trace – if humans perform an inference immediately and only need to verbalize one reasoning step, that step was more likely to involve implicit cues with contextual nuances (e.g., rapidly interpreting body language based on external knowledge). For samples with 1 reasoning step, 53% referenced external knowledge in this first piece of evidence, in contrast to 33% of samples with 5 reasoning steps and 20% of samples with 10 reasoning steps.

We examined model social inference performance across samples with different lengths of human reasoning traces, visualized for $k = 0$ in Figure 9. *Overall, multimodal models social inference performance was lower for samples with shorter human reasoning traces and higher for samples with longer human reasoning traces.* This trend was observed for both larger closed-source models and smaller open-source models that represent different training data distributions, architectures, and training techniques. For example, Gemini-1.5-Flash and LLaVA-Video achieved accuracies of 70% and 58%, respectively, for samples with the shortest reasoning traces and both achieved 80% for samples with the longest reasoning traces. These results reinforce the perspective that current models (regardless of size) are not sufficient for strong multimodal social reasoning performance in domains requiring more contextual understanding.

D.2 Does Chain-of-Thought Prompting Elicit Multimodal Social Reasoning?

Chain-of-thought (CoT) prompting (Wei et al., 2022) has been a prevalent approach to elicit model reasoning abilities in domains such as mathematics

Table 3: Human evaluation mean annotator scores for reasoning trace samples across models.

Model	Shot	Fine-Grained	Compositional	Modality Tags	Validity	Comprehensive
VideoChat2	0.0	1.50	1.00	0.5	0.0	1.25
	1.0	1.00	1.00	1.0	0.0	1.00
	4.0	1.00	1.00	1.0	0.0	1.00
	16.0	1.00	1.00	1.0	0.0	1.00
Video-ChatGPT	0.0	1.00	1.00	1.0	0.0	1.00
	1.0	1.00	1.00	1.0	0.0	1.00
	4.0	1.00	1.00	1.0	0.0	1.00
	16.0	1.00	1.00	1.0	0.0	1.00
LongVA	0.0	2.75	2.75	1.0	1.0	4.50
	1.0	3.00	3.25	1.0	1.0	3.75
	4.0	2.25	3.50	1.0	1.0	3.50
	16.0	2.25	2.75	1.0	1.0	3.25
LLaVA-Video	0.0	3.00	2.50	1.0	1.0	4.00
	1.0	2.00	1.75	1.0	0.5	3.00
	4.0	1.00	1.00	1.0	0.0	1.00
	16.0	1.00	1.00	1.0	0.0	1.00
GPT-4o	0.0	3.50	3.00	1.0	1.0	5.00
	1.0	3.50	3.00	1.0	1.0	5.00
	4.0	4.00	3.50	1.0	1.0	4.50
	16.0	4.00	4.00	1.0	1.0	4.50
Gemini-1.5-Flash	0.0	4.25	4.25	1.0	1.0	4.75
	1.0	4.00	2.50	1.0	1.0	4.75
	4.0	3.75	3.75	1.0	1.0	3.00
	16.0	4.25	4.25	1.0	1.0	4.50

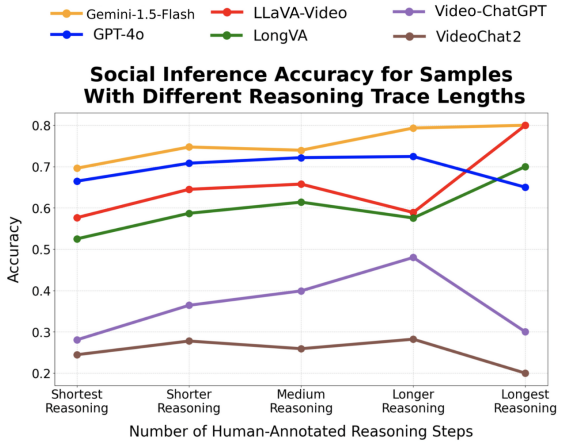

Figure 9: Social inference accuracy of models (k = 0) across samples with different numbers of human-annotated reasoning steps (binned into quintile by reasoning trace length).

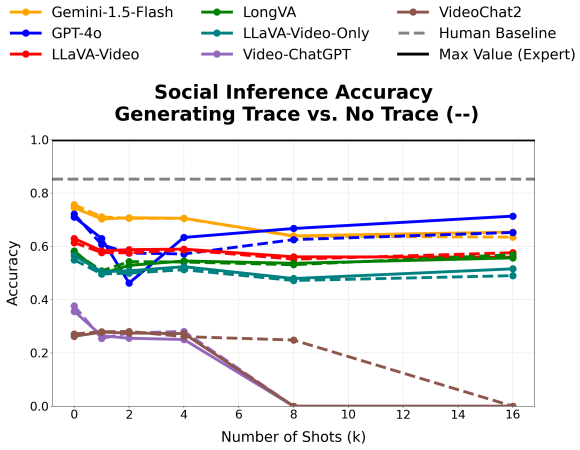

Figure 10: Social inference of models that generated reasoning traces with answers (Trace) compared to models that generated solely answers (No Trace), across different numbers of few-shot samples k.

and code generation (Li et al., 2025). We explore the effectiveness of this technique for multimodal social reasoning with SOCIAL GENOME. Figure 10 visualizes social inference results for models under two settings: *Trace* (models generate step-by-step CoT traces while answering the question) and *No Trace* (models generate only answers and do not generate "step by step" traces). We test models with zero-shot and few-shot settings, with $k \in \{0, 1, 2, 4, 8, 16\}$. Prompts are described in Figure 8.

We find that CoT prompting does not substantially improve the social reasoning performance of models, with the exception of GPT-4o (using CoT performs 6-8% higher than without CoT after k=4). Unlike domains such as mathematics with more formal step-by-step reasoning paths, social reasoning often involves interpreting and integrating ambiguous, context-dependent cues across actors, modalities, and time (Mathur et al., 2024). Social inference is a form of *informal reasoning* that does not often verbalize as a “chain-like" process (Galotti, 1989). Models with CoT prompting have been found to rely on task priors from pretraining distributions (Chochlakis et al., 2024); it is possi-

ble these priors are less effective to elicit social reasoning, compared to other forms of reasoning.

E Human Evaluation Details

Trained annotators analyzed 48 samples from all models at $k \in 0, 1, 4, 16$. The k values and model identities associated with traces were anonymized before presenting samples to annotators in a spreadsheet to rate. For each reasoning trace, annotators watched the corresponding video, read the question, read the answer options, and read the correct answer, in order to rate the quality of the reasoning trace. Human evaluation was conducted to assess reasoning traces along the following dimensions:

Fine-grained: The extent to which reasoning traces were fine-grained was assessed on a scale of 1 to 5, with 1 for no references to low-level behavior and 5 for dense references (e.g., references to low-level behaviors in a majority of steps).

Compositional: Compositionality in reasoning traces was assessed on a scale of 1 to 5, with 1 for minimal compositionality (no cross-references of information across reasoning steps) and 5 for high compositionality across steps.

Comprehensive: The extent to which reasoning traces were comprehensive was assessed on a scale of 1 to 5, with 1 referring to traces missing critical information (not referencing relevant video content) and 5 being fully-comprehensive (capturing all relevant content towards an inference).

Modality Tag Correctness: The accuracy of modality and external knowledge tags for each of the reasoning steps within a given trace was assessed with a binary score (1 for correct, 0 for the presence of any error in the trace).

Validity of Reasoning: The validity of the reasoning in a trace, referring to whether or not the trace represented logical combinations of information. It is possible for a reasoning trace to reference minimal low-level information, yet still represent valid reasoning (motivating the inclusion of this dimension). This dimension was given a binary rating (1 for valid, 0 for invalid).

Raw averages from this human evaluation process are presented in Table 3. We discuss findings from human evaluation in Section 4.2. Annotator agreement was computed with Cohen's Kappa κ (Cohen, 1960). We found strong inter-annotator agreement in ratings across dimensions: $\kappa = 0.87$ for "fine-grained" ratings, $\kappa = 0.92$ for "compositional" ratings, $\kappa = 0.94$ for "comprehensive" ratings, and $\kappa = 0.94$ for "validity of reasoning" ratings. The "modality tag correctness" ratings across samples was 98% with errors specifically occurring in modality tags for VideoChat2 traces.

We note that reasoning trace quality for LLaVA-Video, VideoChat2, and Video-ChatGPT, in particular, decreased at $k = 4$. These findings from human evaluation are aligned with quantitative findings in Section 4 and subjective model generation observations in Appendix C.2.

Our automated metrics yield similar model rankings to human evaluation and strong correlation to human judgements, supporting the reliability of these metrics as proxies for reasoning trace quality. Gemini-1.5-Flash achieves the highest score in human judgements for referencing "Low-Level" behavioral cues and achieves the highest score in automated metrics such as Accuracy, overall semantic similarity (Similarity-Trace), and step-level semantic similarity (Similarity-Num Steps). Similarly, GPT-4o and LLaVA-Video, followed closely by LongVA, score higher than Video-ChatGPT and VideoChat2 on both automated metrics and human judgments. This rank-based alignment indicates that automated proxy metrics can capture reasoning quality signals that human evaluators identified.

These trends observed from rank-based alignment are supported by correlation analyses between model metrics and human judgments. For Gemini-1.5-Flash and LLaVA-Video (highest-performing closed-source model and open-source model), the Similarity-Trace and Similarity-Step metrics exhibit R^2 values of 0.75 and 0.61, respectively, demonstrating strong alignment with human judgments of how effectively models ground reasoning in low-level behavioral cues. Strong correlations indicate that (1) these automated metrics can be viewed as effective predictors of low-level social reasoning quality and (2) there is room for future community research into automated evaluation of social reasoning quality. SOCIAL GENOME contributes a new benchmark and findings for the community to study this direction.

Table 4: Social Inference Accuracy on SOCIAL-GENOME for Models Across Different Numbers of Shots (k)

Model	**Chain Type**	$k = 0$	$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 16$
VideoChat2	Chain	0.2624	0.2779	0.2746	0.2725	0.0000	0.0000
	No Chain	0.2712	0.2793	0.2806	0.2611	0.2483	0.0000
Video-ChatGPT	Chain	0.3560	0.2645	0.2550	0.2503	0.0000	0.0000
	No Chain	0.3762	0.2544	0.2746	0.2806	0.0000	0.0000
LongVA	Chain	0.5828	0.4973	0.5283	0.5451	0.5363	0.5565
	No Chain	0.5767	0.5081	0.5424	0.5410	0.5310	0.5686
LLaVA-Video	Chain	0.6292	0.5828	0.5875	0.5895	0.5606	0.5592
	No Chain	0.6137	0.5767	0.5754	0.5868	0.5518	0.5760
LLaVA-Video-Only	Chain	0.5653	0.5047	0.5081	0.5236	0.4791	0.5155
	No Chain	0.5491	0.4960	0.4993	0.5121	0.4717	0.4899
GPT-4o	Chain	0.7100	0.6292	0.4623	0.6332	0.6669	0.7133
	No Chain	0.7207	0.6063	0.5747	0.5713	0.6252	0.6521
Gemini-1.5-Flash	Chain	0.7443	0.7026	0.7073	0.7052	0.6393	0.6534
	No Chain	0.7564	0.7106	0.7052	0.7046	0.6380	0.6346

Table 5: Model performance for semantic and structural similarity metrics across different numbers of few-shot samples k. These are the raw results (metrics visualized in Figure 3 were normalized, as discussed in the metrics section). * refers to model reasoning traces that were less coherent (Appendix C.2). The highest-performance at each value of k is bolded.

Metric	Model	$k = 0$	$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 16$
Similariy-Trace	VideoChat2	0.4138	0.3954	0.4084	0.4079	0.0734	0.0597
	Video-ChatGPT	0.4484	0.4484	0.4533	0.4354	0.0266	0.0760
	LongVA	0.4533	0.4865	0.4929	**0.4994**	**0.4994**	**0.4998**
	LLaVA-Video	0.4915	0.4193	0.4552	0.4178*	0.4629*	0.4731*
	GPT-4o	0.4631	0.4332	0.3437	0.4363	0.4479	0.4648
	Gemini-1.5-Flash	**0.5157**	**0.5060**	**0.5021**	0.4891	0.4887	0.4850
Similarity-Step	VideoChat	0.3864	0.3804	0.3856	0.3749	0.0909	0.0760
	Video-ChatGPT	**0.4524**	**0.4822**	**0.4836**	**0.4700**	0.0578	0.1059
	LongVA	0.3898	0.4148	0.4185	0.4231	0.4297	0.4310
	LLaVA-Video	0.4462	0.4090	0.4330	0.4109	**0.4641**	**0.4856**
	GPT-4o	0.4016	0.3540	0.2943	0.3574	0.3682	0.3837
	Gemini-1.5-Flash	0.4340	0.4378	0.4311	0.4234	0.4231	0.4210
Similarity-Num Steps	VideoChat2	0.5168	0.4623	0.5983	0.4764	0.0000	0.0000
	Video-ChatGPT	0.3267	0.2053	0.2144	0.1989	0.0000	0.0000
	LongVA	0.4186	0.4529	0.4246	0.4179	0.4164	0.4181
	LLaVA-Video	0.7719	0.6970	**0.9205**	**0.7762***	0.4167*	0.3750*
	GPT-4o	0.6226	0.4480	0.3496	0.4575	0.4440	0.4584
	Gemini-1.5-Flash	**1.0376**	**0.9144**	0.7312	0.6984	**0.6436**	**0.6525**
EmotionMetric	VideoChat2	0.0754	**0.2328**	0.2530	0.2322	0.0034	0.0007
	Video-ChatGPT	0.2358	0.1713	0.1904	0.0702	0.0000	0.0000
	LongVA	0.3345	0.1306	0.3244	0.3075	0.3106	0.3121
	LLaVA-Video	0.3579	0.1077	0.1038	0.2168*	0.1667*	0.2500*
	GPT-4o	**0.6502**	0.2005	**0.6631**	**0.6725**	**0.6577**	**0.6286**
	Gemini-1.5-Flash	0.4959	0.1218	0.4152	0.3782	0.3557	0.3755
DifferenceSequenceMetric	VideoChat	0.2673	0.2690	0.2838	0.2659	0.2141	0.2144
	Video-ChatGPT	0.2508	0.2352	0.2325	0.2335	0.2181	0.2111
	LongVA	0.4266	0.4165	0.4102	0.4202	0.4045	0.3750
	LLaVA-Video	0.3783	0.4819	0.4845	0.4907*	0.2589*	0.1754*
	GPT-4o	0.4582	0.3956	0.3671	0.4167	0.4458	0.4638
	Gemini-1.5-Flash	**0.4591**	**0.5301**	**0.5243**	**0.5312**	**0.5240**	**0.5295**

Table 6: Model Performance for metrics related to step-level evidence from modalities and external knowledge across different numbers of few-shot samples k. These are the raw results (metrics visualized in Figure 3 were normalized, as discussed in the metrics section). * refers to reasoning traces that were less coherent (Appendix C.2). The highest-performance at each value of k is bolded.

Metric	Model	$k = 0$	$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 16$
All Modality Steps	VideoChat2	1.2328	1.2436	1.3573	1.1817	0.6357*	0.6184*
	Video-ChatGPT	0.8526	0.6801	0.6684	0.6638	0.6289*	0.6519*
	LongVA	2.1406	1.7335	1.6669	1.6440	1.5659	1.4525
	LLaVA-Video	1.6148	2.2242	2.0526	1.9231*	0.7500*	0.5625*
	GPT-4o	2.3245	1.7155	1.4566	1.7536	1.8676	1.9750
	Gemini-1.5-Flash	**2.4056**	**2.3806**	**2.3529**	**2.3146**	**2.2811**	**2.2798**
Visual Steps	VideoChat2	0.6097	0.4105	0.5821	0.6326	0.0168*	0.0000*
	Video-ChatGPT	0.2677	0.0919	0.0774	0.0755	0.0027*	0.0000*
	LongVA	1.1299	0.9246	0.8863	0.8883	0.7782	0.7442
	LLaVA-Video	1.0301	0.9367	0.9269	0.9580*	0.0000*	0.0000*
	GPT-4o	**1.5894**	**1.4196**	1.0725	**1.4180**	**1.4922**	**1.5598**
	Gemini-1.5-Flash	1.3276	1.2806	**1.1860**	1.1395	1.0431	1.0513
Verbal Steps	VideoChat2	0.8338	0.8836	0.8358	0.8351	0.6316*	0.6184*
	Video-ChatGPT	0.6533	0.5831	0.5774	0.5777	0.6030*	0.6519*
	LongVA	0.8096	0.5942	0.6528	0.5989	0.6464	0.6988
	LLaVA-Video	0.7643	**0.7880**	0.7487	**0.8042***	**0.9167***	0.6875*
	GPT-4o	0.7249	0.2811	0.3984	0.3057	0.3150	0.3320
	Gemini-1.5-Flash	**0.9316**	0.7728	**0.7709**	0.7695	0.7360	**0.7196**
Vocal Steps	VideoChat2	0.1810	0.1635	0.2066	0.0740	0.0000*	0.0000*
	Video-ChatGPT	0.0578	0.0202	0.0188	0.0117	0.0000*	0.0000*
	LongVA	**0.5162**	0.2288	0.1844	0.1359	0.1365	0.1243
	LLaVA-Video	0.2486	0.2375	0.1269	0.0490*	0.0000*	0.0000*
	GPT-4o	0.4302	0.1270	0.0589	0.1184	0.1270	0.1413
	Gemini-1.5-Flash	0.4357	**0.2984**	**0.2551**	**0.2291**	**0.2497**	**0.2579**
External Knowledge Steps	VideoChat2	0.0229	0.3546	0.3230	0.2402	0.0054*	0.0000*
	Video-ChatGPT	0.0130	0.0000	0.0000	0.0000	0.0232*	0.0000*
	LongVA	0.3156	0.5565	0.5175	0.7174	0.6259	0.5044
	LLaVA-Video	0.0792	1.4661	1.6128	1.5734*	0.0000*	0.0000*
	GPT-4o	0.6199	0.9108	0.7554	1.0402	1.1568	1.2366
	Gemini-1.5-Flash	**0.9856**	**1.4901**	**1.6190**	**1.6881**	**1.6854**	**1.6847**
Num Steps	VideoChat2	4.9455	4.7840	5.8526	4.7638	3.1158	2.8876
	Video-ChatGPT	3.4316	2.8955	2.8985	3.4564	2.8888	3.0443
	LongVA	3.4233	**2.3573**	**2.1245**	**1.9455**	**1.8689**	**2.0475**
	LLaVA-Video	**2.5574**	4.0755	4.7333	4.4336*	2.2500*	2.5625*
	GPT-4o	4.0921	4.0757	3.5935	3.7699	3.7583	3.8039
	Gemini-1.5-Flash	5.1060	4.2197	4.4624	3.9747	4.0445	3.6854