Read the Room: VIDEO SOCIAL REASONING WITH MENTAL-PHYSICAL CAUSAL CHAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

"Read the room," or the ability to infer others' mental states from subtle social cues, is a hallmark of human social intelligence but remains a major challenge for current AI systems. Existing social reasoning datasets are limited in complexity, scale, and coverage of mental states, falling short of the rich causal dynamics found in real-life interactions. In this work, we introduce R^3*-Bench*-an evaluation benchmark with fine-grained annotations of belief, intent, desire, emotion, and their causal chains in complex scenarios; and R^3*-FDT*, a large-scale training set generated through a novel automated pipeline with the same structure. We conduct a comprehensive evaluation of state-of-the-art (SOTA) LVLMs on R^3*-Bench*, revealing substantial gaps in consistent multi-step social reasoning. We also fine-tune a 7B model using group relative policy optimization (GRPO) on R^3*-FDT*, achieving notable improvements across multiple social reasoning benchmarks. Our contributions are three-fold: (i) a novel benchmark with richly annotated, multi-step causal reasoning data; (ii) systematic evidence that SOTA LVLMs fall far short of human-level reasoning; (iii) a scalable training dataset that significantly enhances social reasoning performance. We will release our dataset, code and models upon acceptance.

1 INTRODUCTION

"Read the room" requires employing Theory of Mind (ToM) (Premack & Woodruff, 1978) to read others' minds and perform social reasoning with subtle cues; it represents higher-level social intelligence, and plays a key role in helping people navigate social scenarios smoothly. Humans are innate with the ability to perceive huge hidden information from very simple cues (Heider & Simmel, 1944; Fan et al., 2022; Zhu et al., 2020); however, it remains a great challenge for current AI. As illustrated in Figure 1, the visible physical world is only the tip of the iceberg; beneath it lies a vast and often invisible mental world. In just a few seconds of social interaction, people perceive layers of causally linked mental states: who is aware of what, who is hiding what, and how others respond. These interpretations rely not only on observable actions but also on unspoken norms and contextual reasoning. Effective social reasoning thus involves (i) detecting subtle behavioral cues, (ii) estimating diverse and evolving mental states, and (iii) identifying causal chains that connect the physical and mental worlds over time.

Large language models (LLMs) have recently demonstrated strong reasoning abilities across various domains (Brown, 2020; Wei et al., 2022; Kojima et al., 2022; Bubeck et al., 2023; Wang et al., 2024b). However, they still struggle with complex reasoning tasks such as long-term planning and scientific problem solving (Srivastava et al., 2022; Wang et al., 2024c; Mirzadeh et al., 2024; Glazer et al., 2024). Social reasoning—a crucial subset of complex reasoning—also remains challenging for LLMs (Shapira et al., 2023; He et al., 2023; Gu et al., 2024; Wang et al., 2024a). Critically, language alone is insufficient for modeling social cognition: visual signals are essential for inferring subtle, layered, and often concealed mental states.

To address this, large vision-language models (LVLMs) (Liu et al., 2024; Zhang et al., 2023; Lin et al., 2023b; Team et al., 2023; 2024b; Hurst et al., 2024) have emerged, enabling multimodal understanding. Yet, current benchmarks provide limited evaluation of (i) diverse mental state estimation and (ii) the consistency and depth of social reasoning in complex interactions. Moreover, there remains no large-scale video dataset covering multiple mental states and their causal relationships, hindering

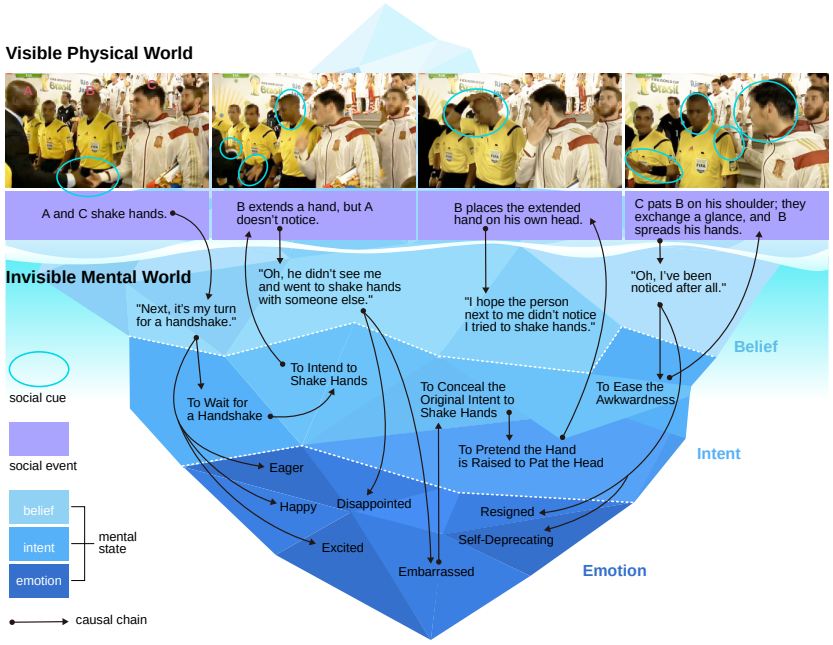


Figure 1: The visible physical world we live in is just the tip of the iceberg compared to the vast, invisible mental world behind it (Zhu et al., 2020). In this example(you, 2014), a brief social interaction reveals complex and dynamic mental activities: B extends his hand to shake with A, but A fails to notice. B then pretends his outstretched hand was meant to touch his head, concealing his embarrassment. C, however, sees through B's mental state and pats him on the shoulder to offer comfort. In response, B shrugs and gestures self-deprecatingly to ease the awkwardness. Social reasoning is a critical aspect of social intelligence. Yet in long-term, dynamic interactions, capturing subtle cues, recognizing social events, estimating mental states, and identifying reasoning chains becomes increasingly difficult, making social reasoning even more intricate.

further development in this area. Current datasets like NExT-QA (Xiao et al., 2021), MVBench (Li et al., 2024), MMBench-Video (Fang et al., 2024), and Video-MME (Fu et al., 2024) focus on factual or visual understanding, offering limited support for mental state reasoning or causal inference. MMToM-QA (Jin et al., 2024b), MELD (Poria et al., 2018), and IntentQA (Li et al., 2023) target mental states like intent or emotion but lack multi-step reasoning. Causal-VidQA (Li et al., 2022) and CausalChaos (Lam et al., 2024) include causal elements but are limited to physical events or rely on animated content. SocialIQ (Zadeh et al., 2019) and Social-IQ 2.0 (Wilf et al., 2023) incorporate social contexts but do not model fine-grained causal chains or assess reasoning across linked events.

We introduce **R**ead-the-**R**oom **R**easoning for **V**ideo **Q**uestion **A**nswering (R^3-*VQA*), a new VideoQA dataset composed of a fine-grained evaluation benchmark (R^3-*Bench*) and a large-scale training set (R^3-*FDT*). It captures rich social interactions and includes: (i) detailed social events, (ii) mental states and their transitions, and (iii) multi-step mental-physical causal chains. We evaluate SOTA LVLMs on R^3-*Bench*, and further fine-tune a 7B model using group relative policy optimization (GRPO) Shao et al. (2024) on R^3-*FDT*. Results show: (i) current models still fall short of human-level social reasoning; (ii) our training data provides notable improvements across several benchmarks.

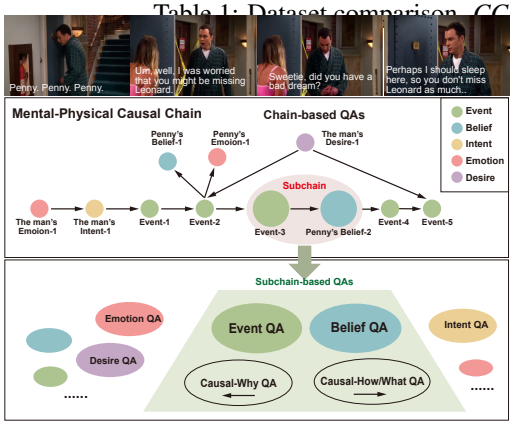
In summary, our contributions are three-fold: (i) we introduce R^3-*Bench*, a novel benchmark with complete, fine-grained annotations for social reasoning; (ii) we show that SOTA LVLMs remain far from human-level performance on this benchmark; (iii) we construct R^3-*FDT* using a scalable automated pipeline, offering valuable training data to improve LVLMs' social reasoning capabilities.

2 R^3-BENCH: A HIGH-QUALITY TESTBED

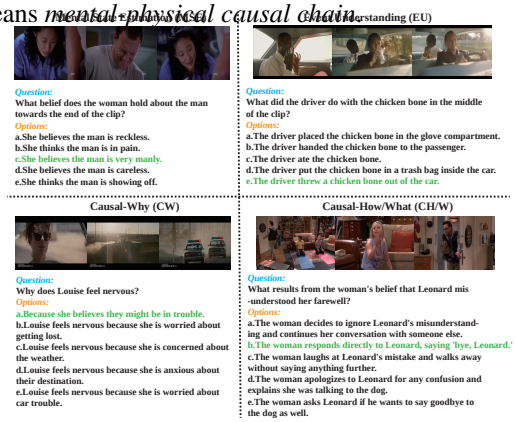
We design our dataset in a natural, intuitive, and principled way, grounded in the foundational Theory of Mind (Premack & Woodruff, 1978), the belief-desire-intention (BDI) framework (Bratman, 1987),

Datasets	Real-World	Mental State				Causality	CC	Training Set		Test Set	
		Belief	Intent	Desire	Emotion			# Video	# MC-QA	# Video	# MC-QA
NExT-QA(Xiao et al., 2021)	✓	✗	✗	✗	✗	✗	✗	3.9k	34k	1k	8.5k
MVBench(Li et al., 2024)	✓	✗	✗	✗	✗	✗	✗	-	-	3.6k	4k
MMBench-Video(Fang et al., 2024)	✓	✗	✗	✗	✓	✗	✗	-	-	0.6k	2k
Video-MME (Fu et al., 2024)	✓	✗	✗	✗	✗	✗	✗	-	-	0.9k	2.7k
MMToM-QA (Jin et al., 2024a)	✗	✓	✓	✗	✗	✗	✗	-	-	0.1k	0.6k
MELD (Poria et al., 2018)	✓	✗	✗	✗	✓	✗	✗	-	-	-	-
IntentQA (Li et al., 2023)	✓	✗	✓	✗	✗	✗	✗	3.2k	12k	0.6k	2.1k
CausalChaos (Lam et al., 2024)	✗	✗	✓	✗	✗	✓	✓	3.4k	3.5k	0.7k	0.7k
SocialIQ [1] (Zadeh et al., 2019)	✓	✓	✓	✓	✓	✓	✗	1k	6k	0.3k	1.5k
Social-IQ 2.0 (Wilf et al., 2023)	✓	✓	✓	✓	✓	✓	✗	1.1k	6.2k	0.3k	1.7k
Our Work											
R³-Bench	✓	✓	✓	✓	✓	✓	✓	-	-	0.3k	5k
R³-FDT	✓	✓	✓	✓	✓	✓	✓	3k	68k	-	-

Table 1: Dataset comparison. *CC* means *mental-physical causal chain*.



(a) We generally illustrate our dataset design (see Section 2.1)



(b) Examples of each QA type. The option marked in green is the correct answer.

Bandura's social cognitive theory—particularly triadic reciprocal determinism (Bandura et al., 1986; Bandura, 1989)—and other modern studies in social cognition (Tomasello, 2010; Pearl, 2014; 2009; Reisenzein, 2006; 2009; Puica & Florea, 2013; Schlaffke et al., 2015; Fan et al., 2022). Our dataset systematically integrates both physical and mental dimensions of social interaction. We include key mental state variables (belief, intention, desire, and emotion) alongside observable physical variables (actions, expressions, dialogue, and other social cues). We further capture comprehensive and dynamic causal interactions among these variables, reflecting the intricate interplay between internal states and external behaviors and environments. This framework positions our dataset as a valuable testbed for developing and evaluating computational models of social reasoning.

2.1 OVERALL DESIGN

R³-Bench differs from traditional VideoQA benchmarks in the following aspects: (i) It includes fine-grained annotations and generation of mental-physical causal chains; (ii) QA pairs are generated based on these causal chains; (iii) It enables comprehensive evaluation of various social reasoning capabilities, including reasoning consistency via causal chains. We illustrate the design and structure of our dataset with an example in Figure 2a.

In our dataset, each video has one or more mental-physical causal chains. We denote a causal chain as $g \in \mathcal{G}$, and a subchain as $g^{sub} \in \mathcal{G}^{sub}$. Here, $\mathcal{G}$ and $\mathcal{G}^{sub}$ respectively represent all the causal chains and subchains. Note that a chain consists of multiple subchains: $g = \{g^{sub}\}$. A subchain g^{sub} comprises one result node v^1 and one or several reason nodes $\{v_i^0\}$. All reason nodes point to the result node via causal edges $\{e_i\}$. Every reason node v_i^0 is necessary to deduce the result node v^1.

For each social causal chain g, we generate a set of related QAs according to the following rules:

(i) For each node $v \in g$, we generate a *Event Understanding* or *Mental State Estimation* QA, depending on its node content. The content of a node is either about an event or a mental state.

[1] Official train/test splits are unavailable. The reported set sizes are estimated using an 80%/20% ratio.

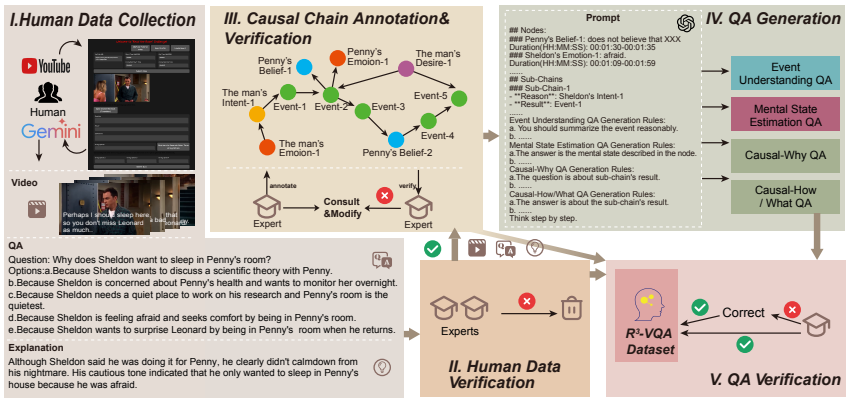


Figure 3: The *R*^{3}*-Bench* construction pipeline, which consists of five stages.

(ii) For each subchain g^{sub}, we generate a *Causal-Why* QA and a *Causal-How/What* QA, depending on whether the reasoning is abductive or deductive.

Therefore, we have four types of QAs in total. We provide four examples selected from our dataset for the four QA types respectively in Figure 2b:

Event Understanding (EU). Event understanding is the premise of social reasoning. We generate a factual QA for each event node. For example, "What happens at the end of the clip?" and "What does Person B do when Person A reaches out her hands?".

Mental State Estimation (MSE). Mental state estimation is a crucial aspect of social intelligence. We consider typical mental states including belief, intent, desire, and emotion. We generate an inferential QA for each mental state node, such as "How does someone feel at the end of the clip?"

Causal-Why (CW) & Causal-How/What (CH/W). *Causal-Why* QAs focus on abductive reasoning and inquire about causes of a given result—for example, the reasons why *Louise feels nervous* in the third case of Figure 2b. In contrast, *Causal-How/What* QAs emphasize deductive reasoning and ask about the effects of a cause, such as the outcome of *the woman's belief* in the fourth case of Figure 2b. To increase question diversity, we include variations such as "How" and "What" in phrasing. An example *Causal-How* QA is: *Question*: "How did Sheldon's revelation affect Penny's emotions?" *Answer*: "Sheldon's comment about the ring's true value led to Penny's disappointment."

2.2 DATASET CONSTRUCTION

R^3*-Bench* is a comprehensive evaluation benchmark featuring high-quality and sufficiently challenging data. This is ensured through three aspects: (i) recruiting participants to independently submit diverse video sources—such as ads, short films, and real-life clips—that meet our criteria; (ii) using LVLMs to filter for difficult and unpolluted samples, meaning data unlikely to appear in standard pretraining sets; and (iii) involving domain experts to verify and annotate the content. To support this, we design a rigorous data construction pipeline illustrated in fig. 3.

Human Data Collection. To support high-quality human-designed QA, we organize a "Read the Room Challenge" to collect data from crowdworkers and volunteers. To guide this process, we establish clear principles and standards: *(1) Data composition:* Each data sample includes a video clip, a question, four distractors, a correct answer, and a reasoning explanation. *(2) Strict standard:* Questions must involve at least one mental state and emphasize causal relations. *(3) No external knowledge required:* Answers must be based solely on observable visual or auditory cues, without relying on external background knowledge. *(4) Explanation requirement:* Each sample must include an explanation describing the reasoning process and underlying causal chain. *(5) Data pollution prevention:* We exclude any sample that Gemini 1.5 Pro answers correctly, ensuring the data remains unseen and challenging. This design ensures that the collected data reflects nuanced, causal, and human-level reasoning that current models are not yet capable of replicating.

Human Data Verification. While many submitted QA pairs successfully challenge Gemini 1.5 Pro, not all meet our quality standards. To ensure consistency, we conduct expert review involving

annotators with backgrounds in cognitive science, linguistics, and AI. Each sample is assessed against our criteria—such as causal depth, mental state relevance, and clarity of explanation—by two independent reviewers. Only those approved by both are retained. Given the nuanced reasoning and social subtleties involved, such validation cannot be reliably automated, highlighting the necessity of human judgment for constructing a benchmark of this complexity.

Causal Chain Annotation & Verification. To capture the reasoning behind each QA, we conduct expert annotation of mental-physical causal chains. Annotators receive training based on strict principles and follow the structured format defined in Section 2.1. Each verified sample is annotated by one of its original reviewers. The annotation process consists of three steps: (i) reviewing the QA and its explanation, (ii) identifying key events and mental states as nodes, and (iii) linking these nodes into subchains according to causal relations. After annotation, the second expert conducts an independent review. If revisions are needed, both experts iterate to reach agreement on the final chain. This multi-stage, fine-grained process ensures that each sample reflects a coherent and interpretable causal reasoning flow grounded in rich multimodal context.

QA Generation & Verification. We use GPT-4o to generate QA pairs for each annotated causal chain, following the rules in Section 2.1. We invite the same experts from the causal chain annotation stage to verify the QA pairs generated from their own annotated chains. Each QA is checked against the following standards: (i) adherence to generation rules, (ii) accurate and unambiguous time references, (iii) full coverage of the corresponding node or subchain content, and (iv) only one correct answer among five options. If a QA fails to meet these criteria, experts revise it following our guidelines or discard it if correction is infeasible. This process yields 4,840 verified QA pairs.

The detailed statistical analysis is provided in Section A.2.2.

3 R^3-*FDT*: AUTOMATICALLY GENERATED DATASET WITH CAUSAL CHAINS

The development of VideoQA datasets for social reasoning is constrained by two significant challenges. First, there is a scarcity of high-quality video content that captures the complexity of social interactions. Second, the annotation process is prohibitively expensive, requiring intensive labor to infer nuanced mental states, followed by extensive verification to reduce ambiguity. These bottlenecks have impeded the development of foundational models in this domain. To overcome these limitations, we introduce an automated pipeline using human annotations and large models, named ARGUS (**A**utomated **R**easoning **G**enerator for **U**niversal **S**ocial Videos). As illustrated in Figure 4, it is designed to generate large-scale training datasets at a reduced cost while maintaining high data quality. Our methodology is centered on the following principles:

(i) Mitigation of Cross-Modal Reasoning Deficiencies: Contemporary models often exhibit limitations such as inconsistencies and hallucinations when processing videos. Our pipeline circumvents these issues by utilizing textual descriptions of videos to enhance the credibility of generated content.

(ii) Assurance of High-Fidelity Annotations: The pipeline exclusively uses human-annotated descriptions that are aligned with videos. It ensures that the foundational data maintains a high degree of accuracy and reliability, consistent with expert annotation standards.

The proposed pipeline offers a cost-effective and scalable solution for generating high-quality training data, addressing a critical need for advancing foundation models in social reasoning.

3.1 INFORMATION ALIGNMENT

To enable large-scale generation of social reasoning data with mental-physical causal chains, we leverage movie data as a rich and structured source. Compared to open-domain videos, movies offer diverse content along with detailed scripts and annotations that describe both visual context and mental states. This makes them naturally suitable for constructing training data that aligns structurally with our human-curated benchmark. We extract information from publicly available movie datasets such as MovieNet (Huang et al., 2020), MovieQA (Tapaswi et al., 2016b), and CondensedMovies (Bain et al., 2020), as well as corresponding movie scripts. As shown in the blue section of fig. 4, we extract key elements from each clip, including scene context, annotated events, mental states, and aligned dialogue. To ensure all generated QAs remain within the temporal

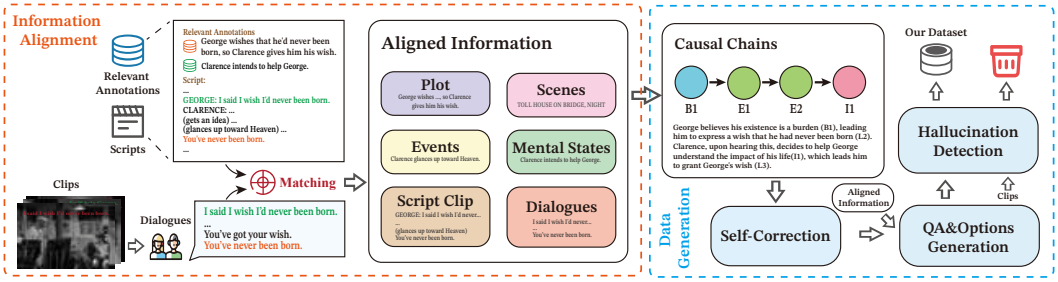


Figure 4: An automated pipeline for generating large-scale video social reasoning data through pure textual data.

boundaries of each clip, we use Whisper (Radford et al., 2023) to align scripts with detected dialogue. The full alignment algorithm is detailed in Section A.3.1.

3.2 DATA GENERATION

Causal Chains Generation. Based on the aligned information, we prompt GPT-4o to infer causal links between events and mental states using contextual cues. In addition to leveraging existing annotations, the model can also identify latent mental states implied by character behavior and dialogue, thereby enriching the reasoning structure with nodes that are not explicitly labeled but contextually grounded. As illustrated in fig. 4, each causal chain is represented symbolically and paired with a natural language explanation. For each clip, the model generates multiple chains that reflect the most salient social interactions. This automated procedure retains the structural depth of human-curated reasoning while making it feasible to build diverse and scalable training data.

Self-Correction To further improve data quality, we require GPT-4o to self-correct the causal chains generated in the previous stage. It needs to check the consistency between the symbolic and textual representations of chains, and remove redundant chains. In this stage, 6% of chains are removed. Although this accounts for only a small proportion, it is necessary to prevent the model from learning from erroneous noises.

QA & Options Generation In accordance with the guidelines outlined in Section 2.1, a QA pair is generated for each node and subchain. To enhance the quality of the data, the inherent causal structure is utilized to create distractors that are both challenging and plausible. Specifically, employing the corrected causal chains and aligned information, GPT-4o generates four distractor options for each question, adhering to two primary principles: (i) the option is designed to be plausible from a common-sense perspective but lacks grounding in the video context; or (ii) the option is grounded in the video context but contradicts the established causal or mentalistic logic. These strategies necessitate a nuanced comprehension of both physical and social cues to arrive at the correct answer.

Hallucination Detection. We establish a hallucination detection stage based to filter for high-quality QA pairs. Specifically, we input the original video and QA pairs into Gemini 2.5 Flash, requiring it to perform a detailed hallucination analysis. During this process, the model is required to complete three key tasks: (i) determine if a QA pair aligns with the video, (ii) provide a detailed explanation, and (iii) output a confidence level. Only when a QA pair is determined to be completely free of hallucination do we retain it in the final dataset. This strict quality control mechanism ensures the reliability of the final dataset, providing high-quality data for training a foundation model.

The resulting training set, *R^3-FDT*, consists of 41k QA pairs distributed across 2,812 videos. This constitutes a large-scale, causally structured dataset. The detailed statistical analysis of the training set is provided in Section A.3.2.

4 EXPERIMENTS

Our experiments consist of two primary components. First, we evaluate a broad set of popular LVLMs on our benchmark, which includes both *R^3-Bench-DX* and *R^3-Bench-Hard*. The *R^3-Bench-Hard* set

is directly sourced from the winning submissions of a social reasoning challenge. We begin by testing several state-of-the-art open-source and closed-source models on the *R³-Bench-DX*, and then select a subset of strong-performing models for further evaluation on the more challenging *R³-Bench-Hard* set. In the second part, we train one of the best-performing models, Qwen2-VL-7B, using GRPO on our training set. We then evaluate the resulting model on our challenge set as well as on external social reasoning datasets to assess its generalization capabilities.

4.1 EVALUATION

Our evaluation approach consists of two dimensions: conventional accuracy metrics and our proposed consistency-based metrics. **(1) Accuracy Metrics.** We report the overall QA accuracy for each model, as well as accuracies across different QA types. In particular, we categorize questions into two main types: *EU* and *MSE*. For the MSE category, we further break it down into four fine-grained subtypes: *Emotion*, *Belief*, *Intent*, and *Desire*. These distinctions help us identify which dimensions of social reasoning are more challenging. **(2) Consistency Metrics.** Despite achieving relatively high QA accuracy, many LVLMs exhibit severe inconsistencies when answering logically related questions. For instance, a model might correctly answer the question "Why did A happen?" with "Because of B," yet fail when asked "Is B present in the video?" shortly after. Humans, in contrast, demonstrate coherent reasoning processes across multi-step chains, maintaining internal consistency throughout.

Table 2: *R³-Bench-Hard* Evaluation Results (%). Other LVLMs' results are illustrated in section A.4.5. All models are given subtitles. "+ Ours-FT" shows results after training on *R³-FDT*.

Model	Overall
Random	20
Idefics3-8B-Llama3(Laurençon et al., 2024a)	24.37
InternVL2-26B(Chen et al., 2023)	24.68
mPLUG-Owl3(Ye et al., 2024)	29.11
InternVL2-76B(Chen et al., 2023)	31.96
GPT-4o Mini[2]	30.70
Gemini 1.5 Pro(Team et al., 2024a) (video)	34.81
Gemini 1.5 Pro(Team et al., 2024a) (frames)	39.24
GPT-4o[3]	48.73
Gemini 2.5 Pro(Team et al., 2024a) (frames)	**59.18**
Qwen2-VL-7B	34.18
Qwen2-VL-7B+ Ours-FT	39.87
Human	**80.06**

To better capture this gap, we introduce two evaluation metrics: **Chain Consistency** and **Subchain Consistency**. These assess whether the model can consistently answer a set of questions derived from a single causal chain or its subcomponents. A chain (or subchain) is marked "consistent" only if the model answers *all* associated questions correctly—partial correctness is not rewarded.

Let $D(g)$ denote the set of all QAs associated with a causal chain g, and $D(g^{sub})$ for subchains g^{sub}. Given a video v and prompt p, the model selects an answer a^* according to:

$$a^* = argmax_{a \in \mathcal{A}} P_\theta(a|v, p) \tag{1}$$

We define **Chain Consistency** as:

$$Cons^c = \frac{\sum_{g \in \mathcal{G}} \prod_{(q,a) \in D(g)} \mathbb{I}(a^* = a)}{|\mathcal{G}|} \tag{2}$$

and **Subchain Consistency** as:

$$Cons^{sc} = \frac{\sum_{g^{sub} \in \mathcal{G}^{sub}} \prod_{(q,a) \in D(g^{sub})} \mathbb{I}(a^* = a)}{|\mathcal{G}^{sub}|} \tag{3}$$

These metrics go beyond surface-level accuracy by evaluating whether a model can sustain coherent reasoning across entire social interaction chains. They are particularly effective at revealing hidden weaknesses in models, where high average accuracy may mask fragmented or inconsistent behavior.

EVALUATION RESULTS AND KEY INSIGHTS

Our evaluation was conducted under two settings: a video-only **basic** setting and a video-plus-subtitle **+Sub** setting. Beyond accuracy on MSE and EU questions, we introduce chain ($Cons^c$) and sub-chain ($Cons^{sc}$) consistency metrics to assess coherent reasoning across related questions.

Two primary insights emerge. First, models find reasoning about mental states (MSE) substantially more challenging than understanding factual events (EU). For example, in the **+Sub** setting, GPT-4o's accuracy drops from 89.77% on EU to 72.44% on MSE, highlighting the difficulty of inferring abstract internal states.

Table 3: ***R³-Bench-DX*** Evaluation Results (%). MSE: mental state estimation. EU: event understanding. CW: causal why. CH: causal how. $Cons^c$: chain consistency. $Cons^c$: sub-chain consistency. while "+ Ours-FT" shows results after training on ***R³-FDT***. The reasons why human consistencies are not reported are elaborated in section A.5.

Model	**Setting**	*MSE*					*EU*	*CW*	*CH/W*	*Overall*	$Cons^c$	$Cons^{sc}$
		Emotion	*Belief*	*Intent*	*Desire*	*Overall*						
Random	-	20	20	20	20	20	20	20	20	20	-	-
Video-LLaVA(Lin et al., 2023a)	-	18.03	18.69	20.22	16.67	18.82	20.26	20.57	21.58	20.33	*0.00*	*0.14*
	+Sub	19.08	19.63	24.38	21.43	20.90	28.28	21.64	22.88	23.14	*0.00*	*0.36*
Idefics2-8B(Laurençon et al., 2024b)	-	10.48	8.41	10.80	2.38	9.74	8.93	8.90	11.48	9.77	*0.00*	*0.07*
	+ Sub	21.80	21.50	22.99	19.05	21.98	20.56	22.99	22.64	22.15	*0.29*	*1.14*
mPLUG-Owl3(Ye et al., 2024)	-	48.01	59.50	42.94	45.24	49.46	52.46	67.90	63.22	58.95	*2.88*	*13.66*
	+ Sub	51.15	71.03	50.97	50.00	56.37	75.03	74.66	70.82	69.21	*9.22*	*25.39*
Phi-3.5-Vision(Abdin et al., 2024)	-	48.85	59.50	49.86	35.71	51.54	53.46	72.31	70.74	62.87	*4.09*	*16.36*
	+ Sub	46.75	62.62	53.46	40.48	52.79	68.71	76.16	72.92	68.00	*8.65*	*23.47*
Idefics3-8B-Llama3(Laurençon et al., 2024a)	-	40.88	57.94	48.75	35.71	47.63	51.15	63.70	62.49	56.82	*3.17*	*12.02*
	+ Sub	45.49	71.65	55.12	54.76	55.70	74.32	74.45	69.44	68.49	*8.07*	*23.83*
PLLaVA-7B(Xu et al., 2024)	-	28.93	26.48	27.98	14.29	27.48	29.29	31.10	32.74	30.25	*0.29*	*1.42*
	+ Sub	22.64	22.43	26.59	23.81	23.81	36.21	29.25	29.02	29.28	*0.00*	*1.21*
PLLaVA-13B(Xu et al., 2024)	-	23.48	22.43	22.71	16.67	22.73	23.77	33.10	30.72	28.00	*0.00*	*0.64*
	+ Sub	27.04	30.84	26.04	26.19	27.73	37.61	38.58	35.57	34.92	*0.58*	*2.56*
PLLaVA-34B(Xu et al., 2024)	-	49.90	59.81	55.40	54.76	54.37	52.46	69.54	70.57	62.52	*5.19*	*16.22*
	+ Sub	53.67	70.40	68.14	69.05	63.03	78.44	77.30	78.42	74.28	*14.12*	*33.50*
InternVL2-8B(Chen et al., 2023)	-	47.17	51.09	42.11	47.62	46.71	46.94	61.78	58.69	54.19	*3.17*	*11.17*
	+ Sub	49.90	68.54	54.57	52.38	56.37	73.22	73.38	68.31	67.83	*8.06*	*25.04*
InternVL2-26B(Chen et al., 2023)	-	42.14	49.22	45.43	45.24	45.13	47.44	59.86	57.56	53.06	*3.17*	*10.60*
	+ Sub	46.12	71.96	58.73	57.14	57.20	73.62	74.45	71.22	69.17	*13.26*	*27.03*
InternVL2-76B(Chen et al., 2023)	-	42.35	63.86	48.48	45.24	50.04	57.87	69.11	66.61	61.43	*5.75*	*15.58*
	+ Sub	53.46	75.70	65.10	64.29	63.28	81.34	79.93	76.39	75.19	*17.29*	*35.99*
GPT-4o Mini[4]	-	43.19	57.63	55.40	52.38	51.04	62.39	73.52	69.36	64.59	6.05	*18.92*
	+ Sub	43.19	57.63	55.40	52.38	51.04	62.39	73.52	69.52	64.63	*6.05*	*18.99*
Gemini 1.5 Flash(Team et al., 2024a) (frame)	-	48.43	65.73	60.94	52.38	56.95	67.60	73.10	70.57	67.31	6.05	*23.33*
	+ Sub	49.90	72.90	65.37	59.52	61.03	80.84	76.87	74.78	73.22	11.24	*33.14*
Gemini 1.5 Pro(Team et al., 2024a) (frame)	-	44.44	65.11	67.31	54.76	57.20	68.61	76.30	73.57	69.28	8.93	*23.97*
	+ Sub	53.67	74.45	77.56	71.43	67.03	85.36	81.71	78.66	78.04	20.75	*44.10*
Gemini 1.5 Flash(Team et al., 2024a) (video)	-	37.95	60.44	53.74	57.14	49.38	69.21	70.39	65.48	63.68	*8.65*	*22.62*
	+ Sub	47.38	72.59	64.27	64.29	59.78	80.34	78.58	73.24	72.91	*11.82*	*30.73*
Gemini 1.5 Pro(Team et al., 2024a) (video)	-	50.31	75.08	74.79	69.05	64.95	84.65	80.13	80.52	77.39	*15.85*	*38.69*
	+ Sub	56.39	74.77	74.79	73.81	67.44	84.55	80.50	77.45	77.31	*19.60*	*41.61*
Gemini 2.5 Pro(Team et al., 2024a) (frame)	-	56.81	83.49	80.06	73.81	77.12	83.85	85.84	81.57	80.79	24.50	*43.53*
	+ Sub	65.83	88.47	85.60	88.10	85.17	93.08	88.33	86.18	86.34	36.60	*58.82*
GPT-4o[5]	-	60.17	80.37	79.50	76.19	71.94	89.17	85.84	82.54	82.23	*25.07*	*47.94*
	+ Sub	61.01	80.69	79.78	76.19	72.44	**89.77**	86.05	82.94	82.64	*25.36*	*48.93*
Qwen2-VL-7B	-	58.07	60.75	55.40	54.76	58.51	59.28	72.31	71.87	65.93	5.48	*19.63*
	+ Sub	58.91	70.40	63.43	59.52	70.02	78.03	79.22	78.66	74.90	12.68	*33.85*
	+ Ours-FT	77.36	83.80	80.33	76.19	83.67	88.16	91.25	87.63	86.88	29.11	*55.83*
Human		**86.67**	**90.00**	**89.19**	**100.00**	**89.09**	**92.04**	90.14	**92.24**	**90.85**	-	-

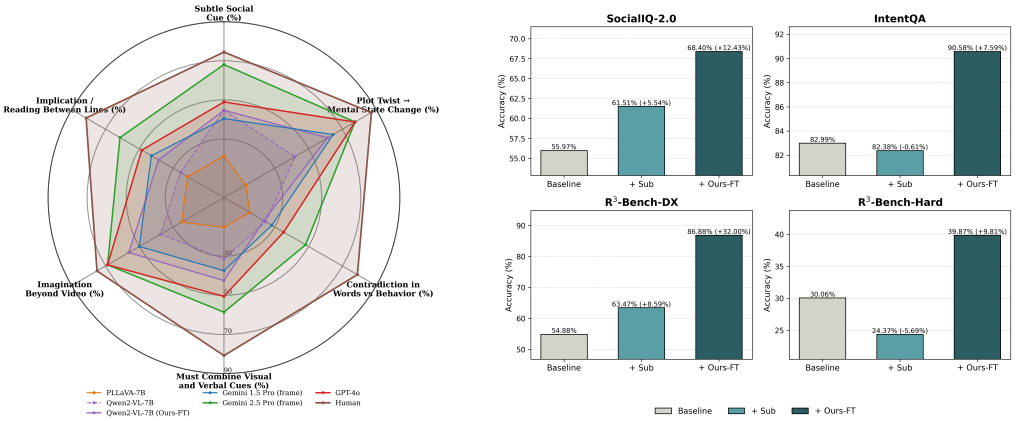
Second, a critical limitation is the models' failure to maintain coherent reasoning, despite high accuracy on individual questions. This disconnect is evident across all models. For instance, in the **+Sub** setting, GPT-4o scores 82.64% in overall accuracy but only 25.36% in chain consistency ($Cons^c$). Gemini 2.5 Pro shows a similar gap, with 86.34% accuracy versus 36.60% consistency. This "high accuracy, low consistency" paradox reveals that models lack a holistic, structured understanding of social interactions, a weakness that single-question accuracy metrics fail to capture.

4.2 Fine-Grained Analysis on Cognitive Dimensions

To delve deeper into the specific reasoning failures of current models, we move beyond a monolithic performance metric to a fine-grained analysis based on six cognitive dimensions. Figure 5a presents a comprehensive performance breakdown of various models across these dimensions, benchmarked against human performance. The results reveal a clear and consistent pattern of deficiencies, even among the most advanced models.

Our analysis yields several key insights. Firstly, the most profound weaknesses are exposed in tasks requiring detection of **Contradiction in Words vs Behavior** and pragmatic inference for **Implication / Reading Between Lines**. Even the best-performing model, Gemini 2.5 Pro, lags significantly behind human accuracy (e.g., 48.2% vs. 78.8% in contradiction detection). This reveals a fundamental deficit in handling modality conflicts and performing pragmatic inference, suggesting models favor literal interpretations over grasping nuanced social contexts involving sarcasm or deceit.

Secondly, the results point to architectural limitations. Current models lack systematic modeling of event structures and evolving psychological states, which explains why even strong models like Gemini 2.5 Pro and GPT-4o struggle with tasks requiring an understanding of mental state



(a) Radar chart of model performance (accuracy %) across six cognitive dimensions. Results highlight the significant gap between SOTA models and human-level reasoning, particularly in tasks requiring detection of multimodal contradictions and pragmatic inference.

(b) Accuracy (%) on four social reasoning benchmarks before and after applying GRPO reinforcement learning. The "Baseline" row reports performance of the pretrained model without fine-tuning, while "+ Ours-FT" shows results after training on *R*^{3}*-FDT*.

Figure 5: Comparison of cognitive and social reasoning capabilities. (a) Model accuracy across six cognitive dimensions. (b) Improvements on social reasoning benchmarks after RLFT.

reversals. In contrast, weaker models such as PLLaVA-7B perform at near-random levels on these tasks, underscoring their inability to transition from mere representation recognition to genuine social-contextual reasoning.

Finally, the **Imagination Beyond Video** category serves as a powerful diagnostic. Here, the performance of top-tier models like Gemini 2.5 Pro and GPT-4o (both at 68.8%) closely approaches the human baseline (75.0%). This highlights the inherent advantage of a powerful linguistic world model for tasks that require reasoning beyond direct perceptual evidence, affirming that strong language priors are crucial for imaginative inference.

In summary, this fine-grained analysis indicates that future advancements in multimodal social reasoning must prioritize: (1) robustly modeling character mental states and motivations; (2) enhancing pragmatic inference for non-literal language understanding; and (3) improving sensitivity to and resolution of contradictions across modalities.

4.3 Training on *R*^{3}*-FDT*

To address previously identified challenges in consistency and multi-step reasoning, we apply reinforcement learning fine-tuning (RLFT) using the GRPO algorithm Shao et al. (2024) on a sampled subset of 13k QA pairs from *R*^{3}*-FDT*, with subtitles incorporated into prompts. Although the training and test sets differ substantially in video domain—our training set consists of curated movie clips, while the test benchmark includes diverse YouTube-style content—they share a unified QA structure grounded in causal chains. This structural alignment allows the model to benefit directly from training signals, leading to performance improvements closely aligned with evaluation objectives. As shown in Figure 5b, training on our dataset yields substantial gains: +32.00% on the *R*^{3}*-Bench-DX* and +9.81% on the *R*^{3}*-Bench-Hard*over the base model.

Further, as reported in Tables 2 and 3, the RLFT-trained Qwen2-VL-7B outperforms GPT-4o on *R*^{3}*-Bench-DX* and surpasses Gemini 1.5 Pro on *R*^{3}*-Bench-Hard*, demonstrating enhanced social reasoning across both QA types. To assess generalization, we also evaluate the model on two external video-based social reasoning datasets: Social-IQ 2.0 and IntentQA. GRPO training brings additional accuracy gains of 6.22% and 5.86%, respectively, confirming the transferability of reasoning capabilities acquired from our dataset.

5 Ethics Statement

Our study involves human participants. Informed consent was obtained from all participants prior to data collection. The released dataset has been carefully anonymized to remove any personally identifiable information, and participants were informed that their data may be shared for research purposes. To mitigate potential misuse, the dataset is distributed under a research-only license, and documentation describing appropriate usage scenarios is provided. We believe that the potential benefits of this dataset for advancing research outweigh possible risks, and we have taken steps to minimize privacy, security, and fairness concerns in accordance with the ICLR Code of Ethics.

6 Reproducibility Statement

We have taken multiple measures to ensure the reproducibility of our results. We provide a detailed description of our data collection method in Section 2 and Section 3, and report implementation details in Section A. To further support reproducibility, we will release our dataset, code, and models upon acceptance.

REFERENCES

Smooth joe fletcher handshake fifa ar spain vs chile, 2014. URL https://www.youtube.com/watch?v=HLcoXVue1Qg. https://www.youtube.com/watch?v=HLcoXVue1Qg.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Albert Bandura. Human agency in social cognitive theory. *American psychologist*, 44(9):1175, 1989.

Albert Bandura et al. Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986(23-28):2, 1986.

Michael Bratman. Intention, plans, and practical reason. 1987.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2), 2022.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

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

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*, 2024.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2023.

Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 709–727. Springer, 2020.

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

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

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Ting En Lam, Yuhan Chen, Elston Tan, Eric Peh, Ruirui Chen, Paritosh Parmar, and Basura Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *arXiv preprint arXiv:2404.01299*, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024a.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21273–21282, 2022.

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

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Mihaela-Alexandra Puica and Adina-Magda Florea. Emotional belief-desire-intention agent model: Previous work and proposed architecture. *International Journal of Advanced Research in Artificial Intelligence*, 2(2):1–8, 2013.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Rainer Reisenzein. *Emotions as metarepresentational states of mind*. na, 2006.

Rainer Reisenzein. Emotional experience in the computational belief–desire theory of emotion. *Emotion Review*, 1(3):214–222, 2009.

Lara Schlaffke, Silke Lissek, Melanie Lenz, Georg Juckel, Thomas Schultz, Martin Tegenthoff, Tobias Schmidt-Wilcke, and Martin Brüne. Shared and nonshared neural networks of cognitive and affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Human brain mapping*, 36(1):29–39, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016a.

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

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

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

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

Michael Tomasello. *Origins of human communication*. MIT press, 2010.

Junqi Wang, Chunhui Zhang, Jiapeng Li, Yuxi Ma, Lixing Niu, Jiaheng Han, Yujia Peng, Yixin Zhu, and Lifeng Fan. Evaluating and modeling social intelligence: A comparative study of human and ai capabilities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024a.

Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Ying Nian Wu, Song-Chun Zhu, and Hangxin Liu. Llmˆ 3: Large language model-based task and motion planning with motion failure reasoning. *arXiv preprint arXiv:2403.11552*, 2024b.

Shu Wang, Lei Ji, Renxi Wang, Wenxiao Zhao, Haokun Liu, Yifan Hou, and Ying Nian Wu. Explore the reasoning capability of llms in the chess testbed. *arXiv preprint arXiv:2411.06655*, 2024c.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

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

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

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

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

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. doi: 10.1609/aaai.v31i1.11238. URL https://ojs.aaai.org/index.php/AAAI/article/view/11238.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020.

Videos	*R³-Bench-DX*					*R³-Bench-Hard* QAs
	EU	MSE	CW	CH/W	Overall	
312	997	1201	1405	1237	4840	316

Chains	Subchains	Nodes					
		Event	Belief	Intent	Desire	Emotion	Overall
347	1406	997	321	361	42	477	2198

Table 4: Statistics of *R³-Bench* dataset

A Appendix

A.1 The Use of Large Language Models (LLMs)

We use LLMs in the following aspects: (1) We utilized LLM in dataset construction process; (2) We evaluated task performance on several LLMs; (3) We finetuned a LLM; (4) We use LLM to assist us with paper writing slightly.

A.2 More Details of *R³-Bench*

A.2.1 Validity of Annotated Causal Chains

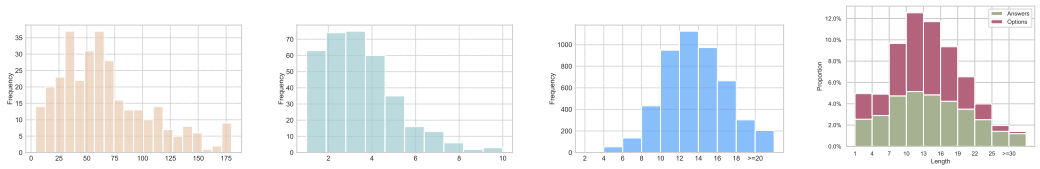
In our dataset, each causal chain is approved by three people: the participant who designed the QA and two experts for verification (one of whom annotated the causal chain). The participant submits the QA and the textual description of the reasoning process behind it, establishing the main content of the causal chain. After being verified by two experts, the filtered data is approved by all three people. The expert in charge of annotation only appropriately expanded and completed the reasoning process without affecting the essence of the causal chain. Also, the annotated causal chains are verified by the other expert. If they fail to pass the verification, the two experts will consult and revise them together. Therefore, we can consider the annotated causal chains to be widely acceptable.

A.2.2 Statistical Analysis

Video Statistics. Our *R³-Bench* dataset contains 312 videos, each annotated with one or more causal chains and multiple QA pairs. Figure 6a shows the distribution of video durations. All videos are under 180 seconds, with an average duration of 66.6 seconds. The increased video length amplifies the challenge of social reasoning, as identifying relevant social cues becomes more difficult, and causal chains may become longer with more steps, involving a greater variety of events and mental states and requiring tracking of all dynamic changes of all states throughout the video.

Causal Chain Statistics. Table 4 presents the statistics of causal chains in our *R³-Bench* dataset. The dataset includes 347 causal chains composed of 2198 nodes and 1406 single-step subchains. These nodes are categorized into 997 *Event* nodes, 321 *Belief* nodes, 361 *Intent* nodes, 42 *Desire* nodes, and 477 *Emotion* nodes. Aside from *Desire*, each mental state category has over 300 nodes, providing ample cases to assess specific mental states. In total, 1201 mental state nodes indicate a rich presence of mental state dynamics. The lengths of causal chains range from 1 to 10 steps, with an average length of 3.3 steps, as shown in Figure 6b. Over 60% have no less than three steps, suggesting that the videos require complex and in-depth reasoning.

QA Statistics. As detailed in Table 4, there are 4840 QAs in *R³-Bench-DX*, and there are 316 QAs in *R³-Bench-Hard*. Figure 6c and Figure 6d show the distributions of question, answer, and incorrect option lengths among *R³-Bench-DX*. The average question length is 13.2 words—longer than many popular VideoQA datasets (Tapaswi et al., 2016a; Zadeh et al., 2019; Fu et al., 2024; Yu et al., 2019; Zeng et al., 2017). Answers and incorrect options average 14.4 and 12.8 words respectively, making the distractors similar in length to the correct answers and thus more challenging.



(a) Video Length (b) Causal Chain Length (c) Question Length (d) *A&O* Lengths

Figure 6: Statistics of *R³-Bench-DX*. *A&O* means *Answer and Options*

A.3 More Details of *R³-FDT*

A.3.1 More Details of Information Extraction

The algorithm to extract annotations from existing movie datasets is shown in algorithm 1. We align the dialogues recognized by Whisper with the time-stamped annotations provided in MovieNet to identify the temporal boundaries of each clip within the full movie. Using these identified temporal boundaries, we then extract the relevant data from both MovieNet and MovieQA, including question–answer pairs, action tags, place tags, and scene descriptions.

Algorithm 1 *ExtractAnnotation*

Require: Video v, Recognized Dialogue D, Movienet Annotation File A, MovieQA Set Q
Ensure: Clip time (t_s, t_e), MovieNet Annotation M, MovieQA Annotation Y

$S_a \leftarrow \texttt{ExtractSubtitle}(A, v)$
$S_d \leftarrow \texttt{ExtractDialogue}(D, v)$
$E_a \leftarrow \texttt{GetEmbeddings}(S_a)$
$E_d \leftarrow \texttt{GetEmbeddings}(S_d)$
$(i_s, i_e) \leftarrow \texttt{MatchIndex}(E_a, E_d)$
if $i_s =$ None **or** $i_e =$ None **then**
    return None
end if
$(t_s, t_e) \leftarrow \texttt{GetTimestamps}(S_a, i_s, i_e)$
$M \leftarrow \texttt{MatchMovieNet}(A, t_s, t_e)$
$Y \leftarrow \texttt{MatchMovieQA}(Q, t_s, t_e)$
return (t_s, t_e), M, Y

Movie scripts include content beyond the target clips. To avoid using such unrelated information, we extract script segments based on aligned dialogues. As shown in algorithm 2, we first identify character names as anchors to locate dialogue in the script. We then parse the script S into two parts: scene descriptions and dialogues, yielding a structured script $S^{'}$. Finally, we match the dialogues recognized by Whisper with those in $S^{'}$ and extract the corresponding script segments aligned with each movie clip.

A.3.2 Statistical Analysis

Video Statistics. Our *R³-FDT* includes 2812 videos. The average video duration is 128.76 seconds and the median duration is 131 seconds. Figure 7a shows the distribution of video durations.

QA Statistics. Figure 7b and Figure 7c show the distributions of question, answer, and incorrect option lengths of *R³-FDT*. The average question length is 9.55, the average answer length is 11.32 and the average option length is 9.88.

[2] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024-07-18
[3] https://openai.com/index/hello-gpt-4o/, 2024-05-13
[4] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024-07-18
[5] https://openai.com/index/hello-gpt-4o/, 2024-05-13

Algorithm 2 *ExtractScript*

Require: Script S, Movie ID I, Recognized Dialogue D
1: $C \leftarrow \texttt{GetCharacters}(I)$
2: $S^{'} \leftarrow \texttt{ParseScript}(S, C)$
3: $S_c \leftarrow \texttt{Grounding}(S^{'}, D)$
4: **if** $\texttt{Len}(S_c) < \epsilon$ **then**
5: **return** S_c
6: **end if**
7: **return** None

Videos	EU	Belief	Desire	Intent	Emotion	CW	Overall
2812	14749	5645	220	1552	4532	14831	41529

Table 5: Statistics of *R³-FDT* dataset

A.4 More Details of Experiments

A.4.1 Configuration

Our evaluation is based on VLMEval[6] repository, an open-source toolkit for assessing LVLMs across multiple benchmarks without extensive data preparation. By extending VLMEval, we incorporates our dataset as a new benchmark.

We process videos according to the models' capabilities. For models that cannot process video data directly (e.g., Idefics2(Chen et al., 2023)), we uniformly sample 16 frames from each video to serve as input. For models that accept video inputs (e.g., Gemini(Team et al., 2024a)), we use raw videos and frames respectively. Additionally, we provide video subtitles generated using Whisper (Radford et al., 2023) and incorporate them into the text prompts.

Our task is formulated as a multiple-choice VideoQA problem. Models receive the raw video or selected frames, along with a question and five options, and must select the correct option. We enforce output formatting constraints and determine selected option through exact matching, ensuring reproducible evaluation results. In Section 4.1, we present our specific evaluation metrics.

A.4.2 GRPO Training Details

We implement GRPO training based on verl[7]. During training, we input 8 frames to Qwen2-VL-7B, each with a resolution of 280×280, and input the dialogue recognized by Whisper to the model. We use a rule-based reward function as following:

$$r_f = \begin{cases} 1 & \text{if the response meets the format requirements} \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

$$r_{acc} = \begin{cases} 2 & \text{if the answer is correct} \\ -2 & \text{otherwise} \end{cases} \tag{5}$$

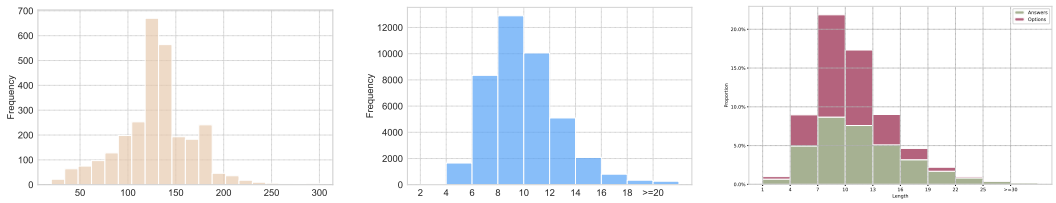
$$r = r_f + r_{acc} \tag{6}$$

We denote the reference model as π_{ref}, the old policy model as $\pi_{\theta_{old}}$ and the policy model as π_θ.

For each question q, we sample a group of responses $\mathbf{o} = \{o_1, o_2, ..., o_N\}$ and compute a group of rewards $\mathbf{r} = \{r_1, r_2, ..., r_N\}$. For each $r_i \in \mathbf{r}$, the corresponding advantage A_i is computed as:

$$A_i = \frac{r_i - mean(\mathbf{r})}{std(\mathbf{r})} \tag{7}$$

[6] https://github.com/open-compass/VLMEvalKit/
[7] https://github.com/volcengine/verl.git



(a) Video Length (b) Question Length (c) *A&O* Lengths

Figure 7: Statistics of *R³-FDT*. *A&O* means *Answer and Options*

Then the policy model π_θ is updated according to the following optimization objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[q \sim P(Q), \{o_i\}_{i=1}^{N} \sim \pi_{\theta_{old}}(O \mid q)\right]$$
$$\frac{1}{N}\sum_{i=1}^{N}\left(\min\left(\frac{\pi_\theta\left(o_i \mid q\right)}{\pi_{\theta_{old}}\left(o_i \mid q\right)}A_i, \text{clip}\left(\frac{\pi_\theta\left(o_i \mid q\right)}{\pi_{\theta_{old}}\left(o_i \mid q\right)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right),$$
$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)} - \log\frac{\pi_{ref}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)} - 1, \tag{8}$$

Both the video context and the dialogue are required to answer *R³-Bench-DX* and *R³-Bench-Hard* in *R³-Bench* and QAs in SocialIQ 2.0 correctly. Therefore, when evaluating on these datasets, we also provide the dialogue to Qwen-2-VL-7B. Previous studies mostly input only the video context to models. When evaluating on IntentQA , we do not provide the dialogue to the model. We input the model with a $16 \times 360 \times 640$ video context when evaluating. We report the training parameters and visualizations in the supplementary material.

A.4.3 Compute Resources

All evaluations were conducted on machines equipped with NVIDIA A100 GPUs. For models with fewer than 10B parameters (e.g., Idefics2-8B (Laurençon et al., 2024b)), we used a single A100 GPU. For mid-scale models in the 20–40B range, such as PLLaVA-34B (Xu et al., 2024) and InternVL2-26B (Chen et al., 2023), we used two A100 GPUs. Large-scale models with over 70B parameters, including InternVL2-76B (Chen et al., 2023), were evaluated using four A100 GPUs. During the training stage, all models were trained using four NVIDIA A100 GPUs.

A.4.4 Additional Results on *R³-Bench-DX*

Further Classification of Subchains We further divide subchains (stand for social causality) into six categories. Category I is MS $\rightarrow$ E, which means that an event is the result of one or more mental states. Category II is MS & E $\rightarrow$ E, which means that an event is the result of the combination of one or more mental states and one or more events. Category III is E $\rightarrow$ E, which means that an event is the result of one or more events . Category IV is E $\rightarrow$ MS, which means that a mental state is one or more events. Category V is E & MS $\rightarrow$ MS, which means that a mental state is the result of the combination of one or more events and one or more mental states. Category VI is MS $\rightarrow$ MS, which means that a mental state is the result of one or more mental states. Through division, we can delve deeper into discussing Large Vision-Language Models' (LVLMs') performance on various types of social causality and analyze their strengths and weaknesses.

We have quantified the number of *CW* QAs and *CH/W* QAs generated according to each of the six categories of subchains, as presented in Table 6. Among these, the questions that involve mutual reasoning between MS and E (i.e., Categories I and IV) are the most prevalent. In contrast, questions that employ both E and MS to reason about MS (i.e., Category V) are the least numerous and constitute the most challenging subset of QAs.

Table 6: Total Quantities per Category. The table categorizes social causality into six categories: I. MS → E. II. MS & E → E. III. E → E. IV. E → MS. V. E & MS → MS. VI. MS → MS.

Category	Total Subchains	*CW*	*CH/W* QAs	Total QAs
I	434	434	392	826
II	199	199	170	369
III	96	96	91	187
IV	464	463	403	863
V	79	79	65	144
VI	134	134	116	250
Total	**1406**	**1405**	**1237**	**2642**

Table 7: Additional Results on *R³-Bench-DX*. The table categorizes social causality into six categories: I. MS → E. II. MS & E → E. III. E → E. IV. E → MS. V. E & MS → MS. VI. MS → MS. **All values are reported as percentages (without % symbols). The $Cons^{sc}$ columns are shown in bold. while "+ Ours-FT" shows results after training on *R³-FDT*.**

Setting	Method	I			II			III			IV			V			VI		
		CW	CH/W	$Cons^{sc}$	CW	CH/W	$Cons^{sc}$	CW	CH/W	$Cons^{sc}$	CW	CH/W	$Cons^{sc}$	CW	CH/W	$Cons^{sc}$	CW	CH/W	$Cons^{sc}$
Video-LLaVA	-	20.51	19.90	**0.00**	22.61	22.35	**0.00**	12.50	9.89	**0.00**	21.81	23.33	**0.22**	22.78	29.23	**0.00**	17.16	25.00	**0.75**
	+ Sub	21.20	22.96	**0.23**	24.12	23.53	**0.00**	13.54	12.09	**0.00**	23.76	23.82	**0.86**	22.78	27.69	**0.00**	16.42	24.14	**0.00**
Idefics2-8B	-	9.68	10.20	**0.00**	10.05	14.12	**0.00**	11.46	12.09	**0.00**	7.56	11.66	**0.00**	8.86	10.77	**0.00**	7.46	11.21	**0.75**
	+ Sub	23.96	23.98	**1.15**	22.11	22.94	**1.01**	28.12	21.98	**1.04**	23.33	21.34	**1.72**	20.25	20.00	**0.00**	17.91	24.14	**0.00**
mPLUG-Owl3	-	59.91	66.07	**13.36**	71.36	65.29	**9.55**	77.08	65.93	**18.75**	73.00	60.05	**17.03**	60.76	66.15	**3.80**	68.66	57.76	**11.19**
	+ Sub	65.67	75.77	**23.04**	76.88	76.47	**24.12**	87.50	80.22	**45.83**	79.05	63.03	**28.66**	77.22	70.77	**13.92**	74.63	65.52	**15.67**
Phi-3.5-Vision	-	65.67	74.23	**14.06**	78.89	74.12	**11.56**	78.12	64.84	**15.62**	74.51	66.75	**20.91**	78.48	75.38	**6.33**	68.66	69.83	**21.64**
	+ Sub	68.66	77.55	**20.05**	82.91	82.94	**19.60**	80.21	64.84	**27.08**	79.05	67.74	**28.66**	78.48	69.23	**15.19**	76.12	68.97	**24.63**
Idefics3-8B-Llama3	-	57.14	64.03	**11.29**	68.34	64.71	**9.05**	77.08	57.14	**12.50**	65.23	60.79	**14.01**	60.76	67.69	**5.06**	64.18	61.21	**15.67**
	+ Sub	67.74	75.51	**22.35**	77.89	79.41	**23.62**	83.33	68.13	**35.42**	76.03	62.53	**24.78**	81.01	64.62	**12.66**	75.37	62.07	**23.88**
PLLaVA-7B	-	31.34	33.16	**1.15**	33.67	32.94	**0.50**	30.21	23.08	**1.04**	31.10	34.00	**2.37**	30.38	40.00	**0.00**	26.87	30.17	**1.49**
	+ Sub	29.03	30.10	**0.69**	28.64	28.82	**0.00**	25.00	20.88	**2.08**	31.10	30.02	**2.16**	31.65	29.23	**1.27**	25.37	28.45	**0.75**
PLLaVA-13B	-	31.80	27.30	**0.23**	33.17	32.94	**0.00**	28.12	21.98	**1.04**	34.77	33.75	**1.08**	37.97	36.92	**0.00**	31.34	31.90	**1.49**
	+ Sub	38.02	33.42	**2.07**	37.19	37.65	**0.50**	40.62	29.67	**5.21**	39.96	37.72	**3.02**	37.97	41.54	**1.27**	35.82	33.62	**4.48**
PLLaVA-34B	-	61.98	71.68	**14.29**	72.86	70.00	**8.54**	80.21	70.33	**16.67**	72.79	70.47	**21.55**	72.15	66.15	**5.06**	68.66	70.69	**21.64**
	+ Sub	73.50	83.16	**33.87**	82.41	83.53	**32.16**	82.29	84.62	**50.00**	80.78	70.97	**33.19**	87.34	72.31	**20.25**	78.36	78.45	**31.34**
InternVL2-8B	-	52.53	56.38	**9.22**	63.32	61.18	**7.04**	73.96	59.34	**15.62**	66.95	60.05	**13.36**	58.23	66.15	**7.59**	64.93	53.45	**14.93**
	+ Sub	64.52	71.68	**22.12**	76.38	79.41	**25.13**	80.21	68.13	**40.62**	78.62	62.28	**26.72**	73.42	67.69	**15.19**	74.63	62.07	**23.13**
InternVL2-26B	-	53.23	56.38	**9.68**	60.30	60.59	**7.54**	69.79	46.15	**8.33**	65.01	59.55	**13.36**	59.49	58.46	**2.53**	55.97	58.62	**14.93**
	+ Sub	66.59	76.28	**28.11**	72.86	77.65	**22.11**	83.33	74.73	**39.58**	79.91	64.52	**28.23**	79.75	64.62	**8.86**	73.88	68.97	**28.36**
InternVL2-76B	-	61.75	69.90	**15.67**	74.37	73.53	**10.05**	75.00	62.64	**14.58**	73.00	62.53	**19.40**	72.15	67.69	**8.86**	65.67	62.07	**14.93**
	+ Sub	73.50	79.85	**36.18**	85.43	88.24	**35.68**	84.38	86.81	**51.04**	82.51	66.75	**36.21**	81.01	70.77	**25.32**	79.85	75.86	**30.60**
GPT-4o mini	-	66.59	71.68	**16.59**	80.90	72.35	**16.08**	86.46	64.84	**27.08**	76.46	68.24	**22.41**	72.15	72.31	**6.33**	66.42	62.93	**20.15**
	+ Sub	66.59	72.19	**16.82**	80.90	72.35	**16.08**	86.46	64.84	**27.08**	76.46	68.24	**22.41**	72.15	72.31	**6.33**	66.42	62.93	**20.15**
Gemini 1.5 flash (frame)	-	67.74	73.47	**21.20**	78.89	75.29	**22.11**	79.17	72.53	**31.25**	74.51	65.76	**24.78**	75.95	73.85	**13.92**	70.90	67.24	**26.87**
	+ Sub	72.35	78.57	**31.57**	78.39	83.53	**34.17**	87.50	81.32	**50.00**	77.97	67.25	**33.41**	75.95	70.77	**13.92**	78.36	72.41	**35.07**
Gemini 1.5 pro (frame)	-	72.35	73.21	**23.96**	79.40	81.18	**18.59**	79.17	75.82	**33.33**	76.67	71.71	**25.86**	81.01	70.77	**11.39**	78.36	69.83	**26.12**
	+ Sub	79.26	82.65	**44.70**	84.92	88.24	**45.23**	85.42	87.91	**58.33**	82.72	70.97	**42.46**	81.01	70.77	**27.85**	78.36	75.00	**45.52**
Gemini 1.5 flash (video)	-	64.06	67.35	**20.51**	69.85	73.53	**20.60**	77.08	68.13	**34.38**	74.30	60.55	**25.22**	75.95	70.77	**16.46**	70.15	59.48	**18.66**
	+ Sub	74.42	77.81	**32.26**	79.90	82.35	**27.14**	86.46	76.92	**43.75**	80.13	64.76	**29.96**	73.42	64.62	**11.39**	82.09	75.86	**35.82**
Gemini 1.5 pro (video)	-	72.58	83.16	**34.79**	81.91	87.06	**35.68**	87.50	81.32	**59.38**	85.31	77.42	**42.24**	84.81	76.92	**25.32**	76.12	74.14	**36.57**
	+ Sub	76.50	80.10	**41.94**	79.90	84.12	**35.68**	86.46	84.62	**50.00**	83.37	71.46	**43.53**	79.75	78.46	**30.38**	79.85	73.28	**43.28**
GPT-4o	-	81.34	85.97	**47.93**	87.94	90.00	**52.76**	89.58	89.01	**64.58**	87.47	76.67	**46.98**	88.61	76.92	**31.65**	87.31	78.45	**41.79**
	+ Sub	81.34	86.48	**49.08**	88.44	90.59	**53.77**	89.58	89.01	**64.58**	87.26	77.42	**48.28**	89.87	75.38	**31.65**	88.81	78.45	**42.54**
Gemini 2.5 pro (frame)	-	82.03	82.14	**43.09**	86.43	85.29	**34.17**	86.46	87.91	**60.42**	87.69	78.16	**44.83**	88.61	80.00	**29.11**	88.81	81.90	**50.75**
	+ Sub	86.41	88.78	**58.53**	89.95	94.12	**61.31**	89.58	96.70	**76.04**	88.55	79.40	**55.17**	88.61	80.00	**48.10**	90.30	84.48	**62.69**
Qwen2-VL-7B	-	63.82	76.53	**17.51**	75.38	75.29	**16.08**	82.29	62.64	**19.79**	76.46	67.74	**21.77**	74.68	72.31	**15.19**	72.39	72.41	**26.87**
	+ Sub	71.43	84.69	**31.11**	82.41	84.71	**31.66**	93.75	79.12	**55.21**	80.99	71.71	**34.91**	77.22	73.85	**18.99**	84.33	75.86	**35.82**
	+ Ours-FT	88.71	91.58	**54.84**	96.48	90.59	**56.28**	95.83	89.01	**68.75**	90.06	83.87	**57.76**	94.94	90.77	**43.04**	90.30	80.17	**50.00**

Results and Analysis We conducted a further analysis of the *R³-Bench-DX* by categorizing them according to the six subchain classifications outlined in Section A.4.4. As shown in Table 7, we employed the same models and settings reported in the main paper to evaluate our dataset. These settings included scenarios without subtitles (-) and with subtitles (+ Sub). The metrics we report are similar to those in the main paper, encompassing both the accuracy of the QAs and the Subchain consistency ($Cons^{sc}$) for each subchain. However, we further disaggregated the performance statistics based on the six subchain categories. This detailed analysis revealed several intriguing insights:

Subchain Consistency for Mental State Reasoning: Reasoning about events generally exhibits higher Subchain consistency compared to reasoning about mental states. Specifically, Categories I, II, and III, which involve inferring events from events or mental states, demonstrate higher performance than Categories IV, V, and VI, which involve inferring mental states from events or mental states. Notably, event-to-event reasoning (Category III) achieves the highest accuracy and Subchain consistency, significantly outperforming the other categories that involve mental state

reasoning (Categories I, II, IV, V, and VI). This indicates that current LVLMs are more adept at factual causal reasoning than at inferring mental states, which remains more challenging.

Difficulty in Cross-Domain Reasoning: Inferring mental states from events or mental states is typically more challenging than inferring events from events or mental states. In Categories I and II, the core of the reasoning process is based on mental states (MS), with events (E) serving as auxiliary information to aid in accurately inferring the events. This auxiliary role of events facilitates correct reasoning, resulting in Categories I (MS inferring E) and II (MS & E inferring E) not exhibiting a pronounced increase in difficulty. Conversely, Category V, where both events (E) and mental states (MS) are used to infer mental states (MS), is the most difficult, significantly more so than Category IV, which involves events (E) inferring mental states (MS). In these categories, the mental states are the primary focus of inference, and the inclusion of additional mental state information introduces complexity that leads to a substantial decline in model performance. This further underscores the current limitations of LVLMs in inferring mental states.

A.4.5 Additional Results on *R^3-Bench-Hard*

We report the accuracies of other LVLMs on *R^3-Bench-Hard*, which is shown in Table 8. We can see that ToM prompting can still improve models' accuracies.

Table 8: Additional Evaluation Results of *R^3-Bench-Hard*.

Model	**Overall**
Random	20
Idefics2-8B	15.19%
Video-LLaVA	18.35%
Phi-3.5-Vision	23.73%
PLLaVA-7B	17.09%
PLLaVA-13B	20.25%
PLLaVA-34B	30.06%
InternVL2-8B	24.68%
Gemini 1.5 Flash (video)	28.48%
Gemini 1.5 Flash (frame)	30.38%

A.5 Human Study

Human studies are completely harmless to human subjects, and we clearly explain all requirements to the subjects before conducting studies.

We sample 43 causal chains and 481 of *R^3-Bench-DX* for the human study. QAs generated from the same causal chain are interrelated. For fairness, we can not provide all QAs generated from the same causal chain with human (however, model experiments are set up like this) due to human's memory. Therefore, we ask each subject to answer only one question from the same reasoning chain. Please note that, in this setting, the reasoning consistencies of human are severely underestimated. We recruit 38 subjects to do human study and results are shown in the main paper.

B Conclusion

In this work, we address the challenge of consistent and multi-step social reasoning in video through the lens of LVLMs. We introduce *R^3-VQA*, a large-scale dataset constructed via an automated pipeline, comprising *R^3-Bench* for evaluation and *R^3-FDT* for model development. The dataset includes fine-grained annotations of social events, mental states, and their causal links. Our benchmark highlights that current state-of-the-art LVLMs still struggle to reason consistently across causal chains. Fine-tuning a 7B model with GRPO on our training set leads to notable gains across multiple social

reasoning benchmarks. At present, our automatic QA generation focuses on single-step questions; extending it to multi-step causal reasoning remains future work. The annotated explanations of causal chains also show promise as training captions but are not yet utilized. We hope this dataset and framework will contribute to building socially intelligent, multimodal systems and inspire further progress in this direction.