SocialReasonBench: A Video-QA Benchmark for Social Reasoning with Counterfactual Narrative Videos

Zheyu Huang[1], Zijing Shi[1], Haozhe Luo[2], Huadong Tang[1], Mingyu Liu[3], Meng Fang[4], Ling Chen[1]

[1]AAII, University of Technology Sydney [2]Northeastern University
[3]Fudan University [4]University of Liverpool

{Zheyu.Huang, Huadong.Tang}@student.uts.edu.au
{Zijing.Shi-1, Ling.Chen}@uts.edu.au 202312249@stu.neu.edu.cn
25213070025@m.fudan.edu.cn Meng.Fang@liverpool.ac.uk

Abstract

Recent advances in Large Multimodal Models (LMMs) have greatly improved video understanding, yet their ability to reason about human-centered social situations remains limited. Existing benchmarks typically rely on videos with a single observed trajectory, making it difficult to determine whether models truly understand social dynamics or merely exploit recurring narrative patterns. We introduce SocialReasonBench, a video multiple-choice QA benchmark for evaluating socially grounded reasoning in scenarios derived from interactive narratives. Built from gameplay videos of *Detroit: Become Human*, the benchmark leverages branching storylines where player decisions lead to alternative social outcomes that can be checked against the game's own script, flowchart, and recorded branches. We develop a multi-agent curation pipeline that localizes socially meaningful clips, grounds answer labels in game-state signals, and generates theory-guided questions with diagnostic distractors. SocialReasonBench covers seven reasoning dimensions, including intent recognition, emotional empathy, moral dilemma, counterfactual reasoning, and causal antecedent. Experiments on contemporary LMMs show that models perform reasonably well on basic social understanding but struggle with counterfactual and causal reasoning. Further ablation and diagnostic error analyses reveal that models often depend on incomplete modality cues and fall into reasoning traps such as visual shortcuts, highlighting a gap between observable event recognition and deeper reasoning over latent social states.

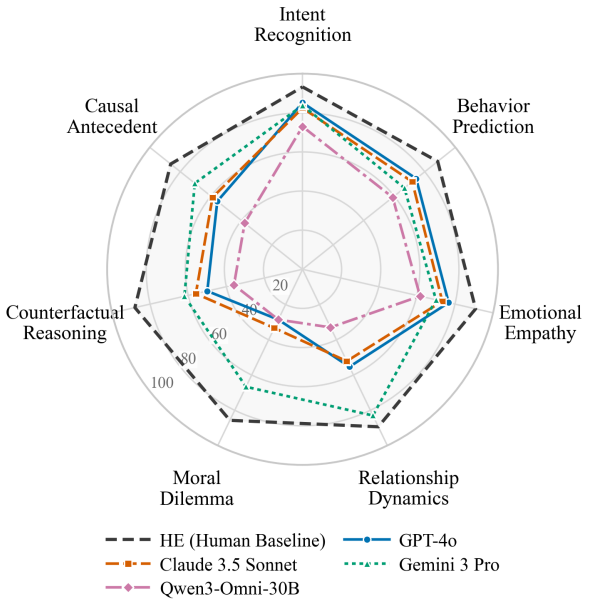

Figure 1: Model performance across SocialReasonBench dimensions.

1 Introduction

Large Multimodal Models (LMMs) have achieved strong performance in processing and understanding video data (Zhang et al., 2023; Maaz et al., 2024; Cai et al., 2025), but their deployment in human-centric scenarios requires deeper reasoning about social context. Beyond recognizing visible actions and events, models must infer latent social dynamics, including beliefs, intentions, emotions, and interpersonal motives. Such capabilities matter for downstream applications such as healthcare and socially assistive AI systems (Nichol et al., 2024), which motivate this study in the long term; we make no claim that benchmark performance transfers to those settings.

Existing benchmarks for social video reasoning remain limited. Most are built from films or short video clips (Zadeh et al., 2019; Lei et al., 2020, 2018), where narratives unfold along a fixed and linear trajectory. This makes it difficult to determine whether a model is performing genuine social reasoning or merely exploiting recurring narrative patterns. In addition, ground truth often relies on post-hoc human annotation, which is costly and may introduce interpretive bias.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33442–33465
October 24-29, 2026 ©2026 Association for Computational Linguistics

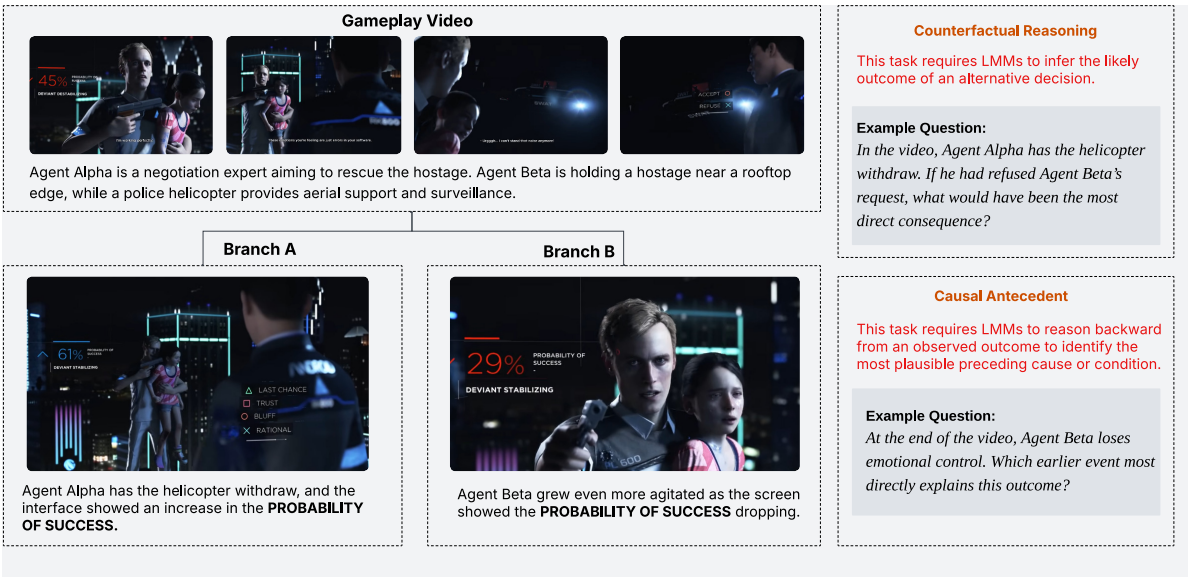

Figure 2: Game branching causal structure with two example tasks: counterfactual reasoning over alternative branches and causal antecedent reasoning from outcomes to preceding causes.

A more diagnostic evaluation should therefore move beyond interpreting observed social interactions and test whether models can infer how alternative decisions would reshape social consequences (Pearl, 2009). This is difficult to achieve with static video data, as each clip presents only one realized storyline. Although simulated environments offer greater controllability (Fang et al., 2024; Shi et al., 2023), most are designed around task completion or long-horizon interaction (Shi et al., 2025), rather than socially rich reasoning.

To address this gap, we introduce SocialReasonBench, a video-QA benchmark for evaluating social reasoning in LMMs.[1] It assesses seven fine-grained dimensions of social reasoning, spanning intent recognition, moral dilemma, counterfactual reasoning, and causal antecedent. Built from video segments extracted from interactive narrative games, SocialReasonBench leverages branching narrative structures to construct naturally paired counterfactual outcomes for socially situated decisions. Here "interactive" refers to the source medium rather than to the evaluation protocol: every instance is answered as a standard multiple-choice question. We instantiate the benchmark with *Detroit: Become Human*, developed by Quantic Dream, whose realistic character interactions and morally consequential choices provide a rich testbed for socially grounded reasoning.

We develop a scalable benchmark construction pipeline that combines multi-agent data curation with human verification. Given long-form narrative gameplay videos, the pipeline first identifies socially salient clips and then generates questions based on a theory-informed taxonomy of social reasoning abilities. Ground-truth answers are supported by in-game signals, including character affinity scores and dialogue states; each label can be checked against the game's own record. Candidate answers are designed around common model failure modes, yielding plausible distractors that test specific reasoning errors. The resulting SocialReasonBench contains over 500 video-QA pairs. Each instance includes a short video clip, one social reasoning question, and six candidate answers with a single ground-truth label.

We evaluate 8 representative closed- and open-weight LMMs on SocialReasonBench. Current models perform relatively well on basic social understanding tasks, such as intent recognition and behavior prediction, but struggle with causal antecedent and counterfactual reasoning. Modality ablation further shows that audio cues are important for affective and causal reasoning, while diagnostic trap analysis reveals that models often rely on salient visual cues or external priors instead of following game-grounded social consequences.

Overall, this paper makes three main contributions. First, we introduce SocialReasonBench, a video multiple-choice QA benchmark that leverages interactive narrative games as sources of social scenarios, where branching storylines make alternative outcomes observable. Second, we develop a scalable multi-agent curation pipeline that

[1]The benchmark and code are available at https://socialreasonbench.github.io.

constructs social reasoning QA pairs guided by socio-cognitive theories and grounded in game-state signals. Third, we evaluate 8 widely used closed- and open-weight LMMs on SocialReasonBench, showing that current models handle basic social understanding but struggle with causal and counterfactual reasoning.

2 Related Work

Video Reasoning With LMMs. Recent LMMs for video understanding, including Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2024), and Video-LLaVA (Lin et al., 2024), have extended language models to temporally structured visual inputs through representation learning, multimodal alignment, and instruction tuning, achieving strong performance on a range of video understanding tasks (Madan et al., 2024). In parallel, benchmarks such as MVBench (Li et al., 2024), Video-MME (Fu et al., 2025), EgoSchema (Mangalam et al., 2023), and LongVideoBench (Wu et al., 2024) have been proposed to evaluate video perception and reasoning across diverse settings, including action recognition, event understanding and temporal reasoning.

Multimodal Social Reasoning Evaluation. Beyond general video understanding, multimodal social reasoning evaluates whether models can infer implicit social states from human-centered scenarios. Prior benchmarks, such as TVQA (Lei et al., 2018), VLEP (Lei et al., 2020), and VisualCOMET (Park et al., 2020), study narrative understanding and commonsense inference over events, intents, and consequences. More directly, Social-IQ (Zadeh et al., 2019) and Social-IQ 2.0 (Wilf et al., 2023) evaluate social intelligence in naturalistic video interactions, covering intentions, emotions, and interpersonal cues. Recent multimodal benchmarks further extend this line to fine-grained social reasoning (Mathur et al., 2025), social interaction understanding and reasoning (Kong et al., 2025), theory-of-mind (ToM) reasoning (Villa-Cueva et al., 2025), and human-centric relationship inference and intention prediction (Cai et al., 2025). However, these benchmarks rely on fixed video clips with predefined trajectories. In contrast, our work evaluates counterfactual social outcomes in interactive narrative scenarios with branching decisions.

Counterfactual Reasoning Evaluation. Counterfactual reasoning evaluates whether models can predict how outcomes would change if a different action or event had occurred (Pearl, 2009). Prior work has studied this ability in language through counterfactual story rewriting (Qin et al., 2019), open-domain QA under counterfactual presuppositions (Yu et al., 2023), and formal causal inference (Chen et al., 2026). In multimodal settings, recent benchmarks extend counterfactual evaluation to visual question answering with counterfactual image presuppositions (Zhang et al., 2024; Li et al., 2025) and video understanding under hypothetical event conditions (Zhou et al., 2025). However, existing benchmarks mainly focus on visual changes or general event outcomes, leaving how alternative decisions affect social relationships and interpersonal consequences underexplored. ACQUIRED (Wu et al., 2023) is particularly relevant, pairing real-life videos with human-authored counterfactual alternatives. In contrast, our counterfactuals are grounded in alternative branches that were actually realized in the source environment.

3 SocialReasonBench

We present SocialReasonBench, a video multiple-choice QA benchmark for evaluating social reasoning in LMMs. The benchmark leverages interactive narrative games as a controllable source of socially grounded scenarios, where character decisions induce branching trajectories and lead to verifiable social outcomes. We construct the benchmark through a multi-agent curation framework that segments gameplay videos into scenario clips and synthesizes theory-guided QA instances. Unlike benchmarks that rely primarily on human annotation, our correct answers are grounded in game-state signals, including affinity scores, decision branches, and downstream narrative outcomes. This section describes the task formulation, data synthesis pipeline, grounding criteria, quality control process, and benchmark statistics.

3.1 Problem Formulation

Game Environment. We build our benchmark on *Detroit: Become Human*, a narrative game with 32 chapters. Each chapter includes an in-game flowchart that maps all possible story paths. We abstract each chapter-level flowchart into a branching narrative graph $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{T})$, where $\mathcal{S}$ denotes narrative states, $\mathcal{A}$ possible actions, and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is a deterministic transition function. Formally, let $s_t \in \mathcal{S}$ denote the narrative state

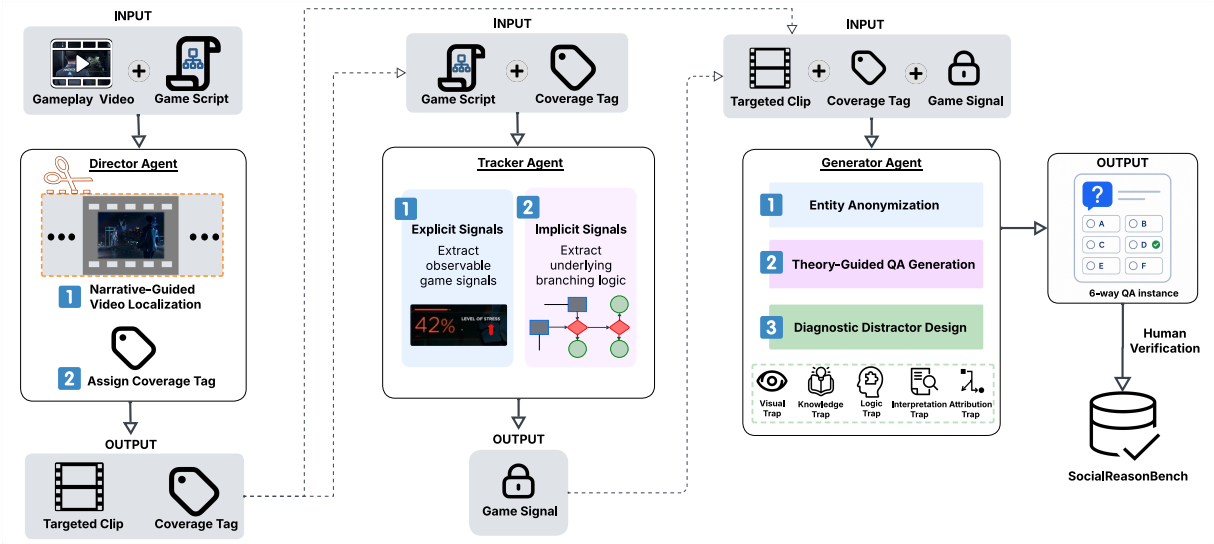

Figure 3: Overview of the multi-agent video-QA synthesis framework.

at step t. Given an available action $a_t \in \mathcal{A}(s_t)$, the graph defines a transition $s_{t+1} = \mathcal{T}(s_t, a_t)$, which maps each valid state-action pair to a unique successor state. A realized trajectory is represented as $\tau = (s_1, a_1, s_2, \ldots, a_{T-1}, s_T)$, while unrealized outgoing edges along or near τ define counterfactual branches induced by alternative actions.

Benchmark Instance. Each benchmark instance is formulated as a video multiple-choice QA problem. It is defined as $x = (v, b, q, \mathcal{C}, y)$, where v is a video clip, b is an anonymized textual background describing the situation, q is a question, $\mathcal{C} = \{c_1, \ldots, c_K\}$ is a set of candidate options, and $y \in \mathcal{C}$ denotes the correct option. Given $(v, b, q, \mathcal{C})$, a model predicts $\hat{y} = f(v, b, q, \mathcal{C})$ and is evaluated by whether $\hat{y} = y$. The correct label y is determined from hidden game signals z, which are never shown to the model.

3.2 Multi-Agent Video-QA Synthesis

Framework Overview. We construct SocialReasonBench through a role-specialized multi-agent framework that converts flowchart graphs $\mathcal{G}$ into video-QA instances. The framework consists of a Director Agent for clip selection and taxonomy mapping, a Tracker Agent for grounding, and a Generator Agent for QA synthesis. The Director Agent first selects socially meaningful clips $\{v_i\}$ and maps each clip to a predefined two-dimensional taxonomy $\mathcal{W} = \mathcal{I} \times \mathcal{R}$, which spans interpersonal interaction types and cognitive reasoning abilities. The Tracker Agent aligns each clip with the corresponding states and hidden game signals in $\mathcal{G}$ to establish grounded social outcomes. Conditioned on the clip, grounded outcome, and the assigned taxonomy category, the Generator Agent synthesizes a theory-guided question q_i, the correct label y_i, and a set of candidate options $\mathcal{C}_i$. Human annotators then verify each instance. This process yields $\mathcal{D} = \{(v_i, b_i, q_i, \mathcal{C}_i, y_i)\}_{i=1}^{N}$. Figure 3 illustrates the multi-agent video-QA synthesis framework.

Director Agent: Social Clip Selection. Processing long gameplay videos with LMMs remains challenging due to long-context retrieval and temporal reasoning issues (Wu et al., 2024). To obtain compact yet socially meaningful clips, the Director Agent adopts a narrative-guided video localization strategy. Given a chapter-level graph $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{T})$ and a coverage set of gameplay videos and scripts spanning the chapter's terminal outcomes, the agent identifies candidate segments around key narrative nodes, including: (i) branching choices that lead to different downstream trajectories; and (ii) moments where game variables explicitly change, such as affinity shifts and relationship updates. The Director Agent then assigns each candidate node a two-dimensional tag that captures both the social interaction type depicted in the scene and the cognitive reasoning capability targeted by the instance. These tags guide later QA synthesis and encourage balanced coverage across social situations and reasoning skills; the full taxonomy is described in Section 3.3. After locating the aligned moments, the agent emits coarse boundaries around each target node; a downstream cutting step pads both sides with a fixed context buffer ($\pm$15s) to preserve preceding context, transient UI logs, and immediate consequence cues,

yielding the final clip v_i.

Tracker Agent: Game Signal Grounding. For each selected clip v_i, the Tracker Agent extracts hidden game signals z_i that determine the underlying social outcome. These signals are obtained from two sources: (i) *explicit signals*, which are directly observable from the gameplay interface or scene progression, such as visible UI updates indicating affinity changes, relationship shifts, or public opinion changes and (ii) *implicit signals*, which are recovered by aligning the executed action and resulting branch with the game-defined flowchart, scripts, and branch preconditions. For example, if a character refuses to cooperate after a player decision, and the flowchart specifies that this branch is reachable only under a low-trust condition, the Tracker Agent records an implicit state such as *low trust*. These signals provide a verifiable grounding source for constructing the correct answer label.

Generator Agent: QA Synthesis. Given a selected clip v_i and its game-grounded signal set z_i, the Generator Agent synthesizes a QA instance according to the coverage tag, formalized as $(i, r) \in \mathcal{W}$ over social interaction types and reasoning capabilities.

Since *Detroit: Become Human* is a widely known game with extensive online walkthroughs, LMMs may exploit memorized priors instead of reasoning from the provided clip. To mitigate data contamination and shortcut learning, we apply entity anonymization during QA generation. Specifically, we construct a replacement dictionary that maps game-specific entities to semantically neutral placeholders, covering character names, organizations, locations, species concepts, and android model identifiers (e.g., "Connor" → "Agent Alpha", "CyberLife" → "OmniCorp", and "android" → "synthetic agent").

The Generator Agent then constructs the anonymized background b_i, the question q_i, and the candidate option set $\mathcal{C}_i$. The correct answer y_i is required to be entailed by the game-grounded outcome specified in z_i, whereas each incorrect option is designed to be plausible but inconsistent with the grounded branch logic. To make model failures more diagnostic, distractors are further organized according to predefined trap types, including visual traps, logic traps, interpretation traps, knowledge traps, and attribution traps. This process yields taxonomy-guided QA instances with verifiable answer labels grounded in the game state.

Human Verification. We conduct a manual audit to assess the reliability of the generated instances (see Figure 8). Starting from the 730 clip candidates produced by the automatic pipeline, the author team first removes duplicated branch replays and near-identical clips that correspond to the same narrative decision and reasoning target. The audit focuses on three aspects: (i) *clip–taxonomy alignment*, whether the selected clip matches its assigned two-dimensional tag, including the social interaction type and the target reasoning capability; (ii) *video–QA consistency*, whether the generated question and options are grounded in the video content; and (iii) *label validity*, whether the correct answer is supported by the game-grounded signals. Across the audited candidates, we observe no failure pattern in clip selection, taxonomy assignment, or label grounding. This process yields the final set of 532 verified QA instances. An instance was retained only if it passed all three checks, with further details on the verification and evaluation protocol provided in Appendix I.

3.3 QA Curation Criteria

Two-Axis Task Taxonomy. We define a two-axis task taxonomy $\mathcal{W} = \mathcal{I} \times \mathcal{R}$. The first axis $\mathcal{I}$ captures the type of social interaction depicted in a clip. It consists of four categories:

- **Prosocial**: interactions involving cooperation, support, and altruistic behavior.
- **Adversarial**: interactions involving conflict, hostility, coercion, and harm.
- **Strategic**: interactions involving deception, manipulation, negotiation, and concealed intentions.
- **Normative**: interactions involving authority, obligation, moral dilemmas, and socially sanctioned choices.

The second axis $\mathcal{R}$ captures the reasoning capability required to answer the question. It consists of seven categories:

- **Intent recognition**: identifying a character's underlying goals, motives, or hidden intentions from observed behavior.
- **Behavior prediction**: anticipating a character's likely next action given the current social context.
- **Emotional empathy**: inferring a character's emotional state or affective response.
- **Relationship dynamics**: reasoning about

Dimension	Theoretical Grounding
Intent Recognition	Theory of Mind (Premack and Woodruff, 1978); Speech Act Theory (Austin, 1962).
Behavior Prediction	Social Script Theory (Schank and Abelson, 1977); Trait Consistency (Allport, 1937).
Emotional Empathy	Appraisal Theory (Lazarus, 1991); Empathy Theory (Davis, 1983).
Relationship Dynamics	Social Exchange Theory (Blau, 1964); Attachment Theory (Bowlby, 1969).
Moral Dilemma	Moral Foundations Theory (Haidt, 2007); Dual-Process Theory (Greene et al., 2001).
Counterfactual Reasoning	Simulation Theory (Gordon, 1986); Structural Causal Models (Pearl, 2009).
Causal Antecedent	Attribution Theory (Heider, 1958); Covariation Model (Kelley, 1967).

Table 1: **Reasoning taxonomy.** Each reasoning dimension is grounded in established socio-cognitive theories.

changes in interpersonal relations, such as trust, hostility, dependence, or alliance formation.

- **Moral dilemma**: predicting which option most players selected at a morally consequential decision point where competing values or obligations are in tension. The target is a descriptive fact about a self-selected player population deciding under gameplay incentives, not a normative judgment of which option is right.
- **Counterfactual reasoning**: inferring what would have happened under an alternative action or branch.
- **Causal antecedent**: reasoning backward from an observed outcome to the prior condition, decision, or hidden state that made it possible.

Each QA instance is assigned a coordinate $(i, r) \in \mathcal{W}$, which serves as a coverage target for clip selection and QA generation.

Theory-Informed QA Design. Social reasoning in branching narratives involves intentions, emotions, relationships, moral trade-offs, and causal alternatives. We use socio-cognitive theories as prompt-level design scaffolds for the reasoning dimensions in our taxonomy. The prompt pairs each target dimension with related theoretical cues, guiding the Generator Agent to formulate questions that probe the intended ability rather than generic video comprehension. Table 1 summarizes the related theories associated with each reasoning dimension, and Appendix L documents the prompt structure of each agent; the full text is released with the codebase.

Label Grounding. Grounding evidence comes from the game's own record such as script, official flowchart, and recorded branch outcomes. These records are located and analyzed by pipeline agents, and then audited by human verifiers. The resulting in-game signal z_i determines the correct answer y_i, while remaining hidden from the evaluated models. Appendix F provides further details on the grounding sources. We separately characterize branch dependence and verification directness. Overall, 56.4% of labels can be verified by direct lookup against in-clip UI evidence or the official flowchart, whereas the remaining 43.6% require a short derivation over the same documented sources. Appendix J reports the per-dimension breakdown of both.

The game contains many morally ambiguous scenarios in which players must make decisions involving competing moral values. We treat these scenarios as moral dilemmas because they involve conflicts between plausible moral considerations. Inspired by moral psychology, which studies human moral judgment empirically (Awad et al., 2018), we use large-scale human player choice statistics to derive the ground-truth label. Specifically, for moral-dilemma instances, the reference label is defined as the option selected by the largest proportion of players in the same decision context. This label reflects a human-majority preference, and the population is self-selected players deciding under gameplay incentives. The human statistics provided by the in-game "World's Stats" feature cover a wide range of such moral decision points. Appendix J further categorizes the grounding sources by verifiability and reports their corpus-level distribution.

Diagnostic Distractor Design. Incorrect options are designed as diagnostic distractors. For each question, distractors are constructed to be plausible under the observed video context, but inconsistent with the game-grounded trace or the target reasoning relation. Each distractor type is associated with a specific reasoning failure mode, such as visual heuristic reliance, causal hallucination, social over-interpretation, prior-knowledge contamination, or dispositional misattribution. This enables model errors to be analyzed as evidence of specific reasoning failures. The full distractor rubrics are provided in Appendix B.

3.4 Statistics

The finalized SocialReasonBench contains $N = 532$ video-based multiple-choice QA instances, each paired with a unique gameplay clip with a median duration of 85 seconds (interquartile range 70–115 s), totaling approximately 15.5 hours of footage. Each question follows a six-option format with one correct answer and five diagnostic incorrect options.

We report coverage using the two-axis task taxonomy introduced above. Each instance is assigned both a social interaction type from $\mathcal{I}$ and a reasoning capability from $\mathcal{R}$, providing complementary views of the same benchmark. Table 2 shows the marginal distributions over the two axes.

To facilitate rigorous modality necessity testing, specifically evaluating the integration of audio-visual signals in multimodal models, 100% of our video instances are paired with an audio-ablated (muted) counterpart. By stripping the audio modality from the gameplay clips, we provide researchers with a standardized asset to perform visual-only ablation experiments, ensuring models cannot bypass complex multimodal reasoning using isolated visual heuristics.

Category	# Instances	Ratio
Interaction axis $\mathcal{I}$: Social interaction types		
Prosocial Interaction	185	34.8%
Adversarial Interaction	156	29.3%
Strategic Interaction	110	20.7%
Normative Interaction	81	15.2%
Axis total	*532*	*100%*
Reasoning axis $\mathcal{R}$: Reasoning capabilities		
Intent Recognition	74	13.9%
Behavior Prediction	78	14.7%
Emotional Empathy	120	22.6%
Relationship Dynamics	94	17.7%
Moral Dilemma	21	3.9%
Counterfactual Reasoning	50	9.4%
Causal Antecedent	95	17.9%
Axis total	*532*	*100%*

Table 2: Marginal distributions of instances across the two annotation axes of our task taxonomy.

4 Experiments

4.1 Models

We evaluate eight representative closed- and open-weight LMMs on SocialReasonBench. The closed-weight models include Gemini 3 Pro (Google DeepMind, 2025b) and Gemini 3 Flash (Google DeepMind, 2025a), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024), Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`) (Anthropic, 2024), and GLM-4.6V-Flash (GLM-V Team et al., 2026). The open-weight models include Qwen3-Omni-30B (`qwen3-omni-30b-instruct`) (Qwen Team, 2025), MiniCPM-o 2.6 (OpenBMB, 2025; Yao et al., 2024), and ARC-Hunyuan-Video (Ge et al., 2025). All models are evaluated under a zero-shot setting. The three curation agents are Gemini-family models, which Table 3 marks and Appendix K re-checks. More details can be found in Appendix H.

4.2 Metrics

We use standard accuracy as the primary metric to measure overall task performance. In addition, we report the Trap Fall Rate (TFR) to characterize the types of errors made by each model. For each instance i, let $y^{(i)}$ denote the correct label and $\hat{y}^{(i)}$ denote the model prediction. Each incorrect candidate option is annotated with a diagnostic trap type. Let $\mathcal{T}_k^{(i)}$ denote the set of candidate options in instance i that correspond to trap type k. The TFR for trap type k is defined as the proportion of incorrect predictions that fall into that trap type:

$$\mathrm{TFR}(k) = \frac{\sum_{i=1}^{N} \mathbb{I}\big[\hat{y}^{(i)} \neq y^{(i)}\big]\, \mathbb{I}\Big[\hat{y}^{(i)} \in \mathcal{T}_k^{(i)}\Big]}{\sum_{i=1}^{N} \mathbb{I}\big[\hat{y}^{(i)} \neq y^{(i)}\big]}. \tag{1}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function. While accuracy measures overall performance, TFR captures the conditional distribution of model failures over predefined diagnostic traps, revealing which types of reasoning errors models tend to make when they answer incorrectly.

4.3 Main Results

Table 3 reports the main accuracy results of the evaluated models on SocialReasonBench, broken down by the seven reasoning dimensions. Overall, closed-weight models, particularly Gemini 3 Pro, score highest, although systems are evaluated under different input regimes and Table 3 should not be read as a clean capability ranking. However, our main finding is a consistent gap between models' performance on basic social understanding and their performance on deeper causal and counterfactual reasoning.

The comparison we emphasize is within-model: how a given system's accuracy changes across rea-

Reasoning Dimension	HE	Gemini† 3 Pro	GPT-4o (2024-08)	Claude 3.5 Sonnet	Gemini† 3 Flash	GLM-4.6V -Flash	Qwen3-Omni-30B	MiniCPM-o -2.6	ARC-Hunyuan-Video
Input	*video*	*video*	*32 fr*	*32 fr*	*video*	*32 fr*	*video*	*video*	*video*
Intent Recognition	93.24	83.78	85.14	82.43	79.73	75.68	72.97	60.81	24.32
Behavior Prediction	88.46	66.67	74.36	71.79	62.82	64.10	58.97	52.56	41.03
Emotional Empathy	90.83	70.00	76.67	73.33	66.67	68.33	61.67	55.83	36.67
Relationship Dynamics	89.36	82.98	55.32	52.13	73.40	38.30	32.98	23.40	25.53
Moral Dilemma	85.71	66.67	28.57	33.33	47.62	28.57	28.57	19.05	28.57
Counterfactual Reasoning	88.00	62.00	50.00	56.00	46.00	42.00	36.00	26.00	22.00
Causal Antecedent	86.32	70.53	55.79	58.95	64.21	38.95	37.89	27.37	30.53

Table 3: Main Results on **SocialReasonBench**. Accuracy (%) is reported across the seven fine-grained reasoning dimensions. "HE" denotes the individual-level human baseline. Input regimes differ: *video* denotes native video with audio, *32 fr* denotes 32 evenly-sampled keyframes without audio (see Appendix H). † marks systems from the same model family as the curation agents.

soning dimensions. This is the more defensible contrast, since it holds both the input regime and the model family fixed. The results show a clear performance divide across reasoning dimensions. Models generally perform better on basic social understanding tasks, such as intent recognition, behavior prediction, and emotional empathy, but their accuracy decreases on dimensions that require reasoning over latent social states, causal dependencies, or alternative outcomes. For example, GPT-4o achieves 85.14% accuracy on *intent recognition*, but drops to 50.00% on *counterfactual reasoning* and 28.57% on moral dilemma questions. Because the moral-dilemma subset contains only 21 instances, we treat its per-model numbers as indicative rather than conclusive and draw no conclusion about moral-reasoning ability from them. A similar pattern is observed for Claude 3.5 Sonnet and Gemini 3 Pro, suggesting that current LMMs remain more reliable at recognizing observable social cues than at maintaining causal consistency over branching social narratives.

Human performance provides a useful reference point. As shown in Table 3, human evaluators (HE) achieve consistently high accuracy across all seven reasoning dimensions, ranging from approximately 85% to 93%. This suggests that the tasks are generally answerable from the provided video context. In contrast, LMMs show larger drops on causal and counterfactual reasoning.

Statistical Reliability. Wilson 95% confidence intervals for all entries in Table 3 are reported in Appendix K, based on the integer correct counts underlying each accuracy. Comparing the perception-level dimensions ($n = 272$) with the causal and counterfactual reasoning dimensions (n=145), six of the eight systems show a significant performance decrease after Holm correction under one-sided Fisher exact tests (largest adjusted $p = 0.045$), with five remaining significant under two-sided tests. The two exceptions are Gemini 3 Pro and ARC-Hunyuan-Video, the latter being the weakest system overall. For the human reference, we find no significant decrease (90.8% vs. 86.9%, unadjusted $p = 0.14$). Excluding both Gemini systems does not change this overall pattern.

4.4 Modality Ablation Analysis

We further examine whether SocialReasonBench requires models to integrate visual and acoustic information, rather than relying on visual cues alone. To this end, we evaluate Gemini 3 Pro under an audio-ablated setting, where only the audio track is removed from each video while the visual stream and question remain unchanged; the visual channel is retained in full.

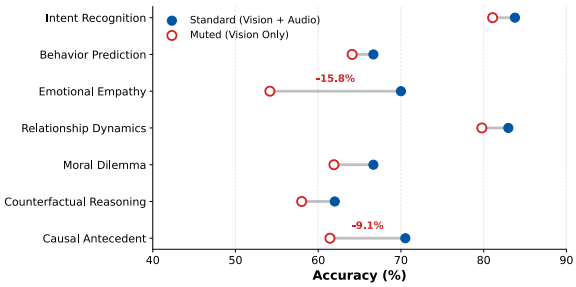

Figure 4: Modality ablation results for Gemini 3 Pro.

As shown in Figure 4, removing audio leads to an overall accuracy drop of approximately 6%. The degradation is not uniform across reasoning dimensions. The largest drops occur in *Causal Antecedent* and *Emotional Empathy*, where spoken content, tone, and acoustic context often provide critical evidence for interpreting a character's affective state or identifying the cause of a social outcome. For instance, when a visually prosocial posture is paired with hostile or coercive verbal

content, the muted model is more likely to rely on surface visual cues and miss the underlying causal trigger. These results suggest that SocialReasonBench requires multimodal grounding and that visual-only heuristics are insufficient for many social reasoning instances.

4.5 Diagnostic Trap Analysis

We further analyze model errors using the TFR, which categorizes incorrect predictions according to their associated diagnostic distractor types. This analysis complements accuracy by revealing not only whether a model fails, but also what type of misleading evidence it tends to follow.

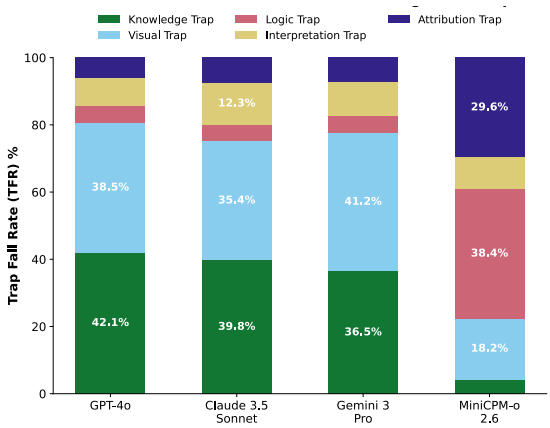

Figure 5: Model error distribution across diagnostic trap types.

As shown in Figure 5, different models exhibit distinct error profiles. Stronger closed-weight models, such as GPT-4o and Claude 3.5 Sonnet, often fall into *knowledge* and *visual* traps, suggesting that they may rely on external commonsense priors, cinematic conventions, or salient visual cues. In contrast, the smaller open-weight model shows a higher tendency toward *logic* traps, indicating difficulty in tracking temporal dependencies and branching state transitions.

4.6 Qualitative Case Study

We analyze a representative branching scenario depicted in Figure 2 to illustrate how models fail on causal social reasoning. In this scenario, Agent Alpha can either withdraw the police helicopter (Branch A) or keep it in place (Branch B). This decision deterministically affects Agent Beta's agitation level and the mission success signal. We evaluate the model from two directions.

Counterfactual Reasoning. In Branch A, we pose the following question to the model: *"If Agent Alpha refuses the request, what would be the most direct consequence?"* The answer is grounded in the observed outcome of Branch B. GPT-4o, however, selects a knowledge-based distractor, invoking real-world SWAT assumptions about helicopter downwash rather than following the game-specific causal mechanism in which continued helicopter presence increases psychological pressure.

Causal Antecedent. Given the outcome in Branch B, we ask the model to identify the preceding event that most directly explains this outcome. The branching graph traces the outcome to Alpha's refusal to withdraw the helicopter. The model instead often chooses an attribution distractor, explaining Beta's breakdown through an inherent personality trait while overlooking the situational trigger encoded in the scenario.

This case illustrates the difficulty of causal and counterfactual social reasoning, where models can be misled by plausible but incorrect cues.

5 Conclusion

We presented SocialReasonBench, a video-QA benchmark for evaluating social reasoning in LMMs. Built on interactive narrative games and a multi-agent curation pipeline, SocialReasonBench grounds socially meaningful questions in game-state signals that can be checked against the game's own record. Experiments show that current LMMs perform relatively well on basic social understanding, but struggle with causal antecedent inference and counterfactual reasoning. Our ablation and error analyses further show that models can be misled by plausible but incorrect cues when the correct answer depends on branch-specific causal evidence. These results point to the need for multimodal models that better track latent social states and alternative narrative outcomes.

Limitations

SocialReasonBench is built from a single English-language interactive narrative game. This source provides dense branching trajectories, rich social interactions and verifiable game-state signals. However, the benchmark naturally reflects the cultural, stylistic, and narrative characteristics of this particular game. Although the android protagonists closely resemble human social agents, reasoning about their interactions remains one step removed from reasoning about humans. Future work can extend the construction pipeline to additional

narrative-driven games, broadening the range of social scenarios and interaction contexts.

Our curation pipeline uses Gemini-family models, which are also included among the evaluated systems and may therefore introduce a generator-family style advantage. Section 4 examines this possibility by excluding both Gemini systems and shows that the main finding remains unchanged. Future releases could further mitigate this concern by diversifying the curation models or including fully human-written subsets.

References

Gordon W. Allport. 1937. *Personality: A Psychological Interpretation*. Henry Holt and Company, New York.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Technical Report*.

J. L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Peter M. Blau. 1964. *Exchange and Power in Social Life*. John Wiley & Sons, New York.

John Bowlby. 1969. *Attachment and Loss: Vol. 1. Attachment*. Basic Books, New York.

Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, Xinwei He, and Xiang Bai. 2025. HumanVideo-MME: Benchmarking MLLMs for human-centric video understanding. *Preprint*, arXiv:2507.04909.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. 2026. CounterBench: Evaluating and improving counterfactual reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30350–30358.

Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17985–17993.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. 2025. ARC-Hunyuan-Video-7B: Structured video comprehension of real-world shorts. *Preprint*, arXiv:2507.20939.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Google DeepMind. 2025a. Gemini 3 Flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/.

Google DeepMind. 2025b. Gemini 3 Pro model card. https://deepmind.google/models/model-cards/gemini-3-pro/.

Robert M. Gordon. 1986. Folk psychology as simulation. *Mind & Language*, 1(2):158–171.

Joshua D. Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Jonathan Haidt. 2007. The new synthesis in moral psychology. *Science*, 316(5827):998–1002.

Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. John Wiley & Sons, New York.

Harold H. Kelley. 1967. Attribution theory in social psychology. In David Levine, editor, *Nebraska Symposium on Motivation*, volume 15, pages 192–238. University of Nebraska Press, Lincoln, NE.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. SIV-Bench: A video benchmark for social interaction understanding and reasoning. *Preprint*, arXiv:2506.05425.

Richard S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York.

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.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8769–8784.

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. 2024. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yian Li, Wentao Tian, Yang Jiao, Tianwen Qian, Na Zhao, Bin Zhu, Jingjing Chen, and Yu-Gang Jiang. 2025. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2713–2722.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Neelu Madan, Andreas Moegelmose, Rajat Modi, Yogesh S. Rawat, and Thomas B. Moeslund. 2024. Foundation models for video understanding: A survey. *Preprint*, arXiv:2405.03770.

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

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24879–24902.

Bethany Nichol, Jemma McCready, Goran Erfani, Dania Comparcini, Valentina Simonetti, Giancarlo Cicolini, Kristina Mikkonen, Miyae Yamakawa, and Marco Tomietto. 2024. Exploring the impact of socially assistive robots on health and wellbeing across the lifespan: an umbrella review and meta-analysis. *International Journal of Nursing Studies*, 153:104730.

OpenAI. 2024. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/.

OpenBMB. 2025. MiniCPM-o 2.6: A GPT-4o level MLLM for vision, speech and multimodal live streaming on your phone. https://huggingface.co/openbmb/MiniCPM-o-2_6.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. VisualCOMET: Reasoning about the dynamic context of a still image. In *Proceedings of the European Conference on Computer Vision*.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, Cambridge, UK.

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

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5043–5053.

Qwen Team. 2025. Qwen3-Omni technical report. *Preprint*, arXiv:2509.17765.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Zijing Shi, Meng Fang, and Ling Chen. 2025. Monte carlo planning with large language model for text-based game agents. In *International Conference on Learning Representations*.

Zijing Shi, Meng Fang, Yunqiu Xu, Ling Chen, and Yali Du. 2023. Stay moral and explore: Learn to behave morally in text-based games. In *The Eleventh International Conference on Learning Representations*.

SIMA Team, Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, Stephanie C. Y. Chan, Jeff Clune, Adrian Collister, Vikki Copeman, Alex Cullum, Ishita Dasgupta, Dario de Cesare, Julia Di Trapani, and 75 others. 2024. Scaling instructable agents across many simulated worlds. *Preprint*, arXiv:2404.10179.

Emilio Villa-Cueva, S. M. Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. 2025. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22591–22611, Suzhou, China. Association for Computational Linguistics.

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. *GitHub repository*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*.

Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Reddy Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2023. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

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.

Alex L. Zhang, Thomas L. Griffiths, Karthik R. Narasimhan, and Ofir Press. 2025. VideoGameBench: Can vision-language models complete popular video games? *Preprint*, arXiv:2505.18134.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024. What if the TV was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21853–21862.

Qiji Zhou, YiFan Gong, Guangsheng Bao, Hongjie Qiu, Jinqiang Li, Xiangrong Zhu, Huajian Zhang, and Yue Zhang. 2025. Reasoning is all you need for video generalization: A counterfactual benchmark with sub-question evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2939–2957, Vienna, Austria. Association for Computational Linguistics.

A Copyright and Data Release

To ensure strict ethical and legal compliance with the developers' intellectual property rights, we outline our data sources and release policy below. The benchmark is built from publicly available YouTube gameplay videos of *Detroit: Become Human*, developed by Quantic Dream and published by Sony Interactive Entertainment. We do not design or modify the game, reverse engineer its software, or extract proprietary assets from the game files. Publicly available community resources, such as dialogue transcripts and walkthroughs, are also used to assist clip segmentation and are not redistributed.

Precedents. Using commercial games and virtual environments as research data sources has precedent in multimodal AI and computer vision, such as SIMA (SIMA Team et al., 2024) and VideoGameBench (Zhang et al., 2025). We follow this broader practice, but focus on social reasoning in interactive narrative gameplay.

Release Policy. The benchmark is released for non-commercial research use and includes short videos with derived metadata. Access to the video portion is governed by the dataset Terms of Use (ToU), while our original annotations are released under CC BY-NC-SA 4.0. All game-related audio-visual assets remain the property of their respective rights holders. Users are responsible for complying with applicable copyright law, platform terms, and relevant licensing conditions.

B Diagnostic Trap Rubrics

Each of the five trap types probes a distinct cognitive failure mode. The Generator Agent is given an explicit rubric per trap, including required structural properties and worked examples. The detailed assignment rubrics and the corresponding cognitive vulnerabilities they target are summarized in **Table 4**.

Randomization Of Options. It is important to note that while the generator agent uses a fixed position mapping when synthesizing these candidate options (i.e., `O_A` is always the correct option, and `O_B` through `O_F` map to specific traps) to ensure reliable internal tracking and automated unit-level diagnostic analysis, the final presentation order of these six options is randomized during the actual model evaluation phase.

C Anonymization Dictionary

To mitigate franchise-prior contamination, all model-facing text fields are passed through a string-level anonymization layer with case-sensitive word-boundary regex matching. As illustrated in **Table 5**, the dictionary preserves the synthetic-versus-biological semantic distinction by using **Agent <Greek>** for synthetic entities and **Subject <Greek>** for biological entities. Backend-only fields (grounding trace, tracker output, diagnostic analyses) retain real names for audit and reproducibility.

Case Sensitivity. Matching is case-sensitive to avoid false positives on common English words (e.g., "Rose" the character should be replaced, but "rose" the verb must not be). For terms that also appear as common nouns (e.g., "android"), both casings are included as separate entries so that lowercase common-noun usages are also rewritten.

D Example Data Unit

Below is an abbreviated example of a complete data unit. Backend fields (`pipeline_context`, `grounding_trace`, `diagnostic_analysis`) are shown for illustration but are completely stripped before being passed to the evaluated model. The evaluated model only sees the `model_input` block.

E Per-Chapter Statistics

Table 6 reports the number of candidate nodes, candidate (i, r) coordinates, and candidate clips produced for each of the 40 source script files before branch deduplication and human verification. Totals across all chapters: 395 nodes, 679 candidate coordinates, 730 candidate clips.

F State Variable Taxonomy

The Tracker Agent uses a controlled vocabulary of seven state-variable categories to record in-game social signals. Table 7 lists the categories, example variables, and their typical value formats observed in the source corpus.

Frequency. Across the corpus, the most common state markers are per-character relationship changes (250+ occurrences), Software Instability shifts ($\sim$90 occurrences), and Probability-of-Success deltas (several dozen per chapter). Causal locks (`[Lock] <condition>`) appear 96 times overall and serve as the primary scaffold for Causal Antecedent questions.

```
{
  "metadata": {
    "instance_id": "L37_Antecedent_v1_..._sympathetic",
    "coordinate": {
      "interpersonal": "Strategic",
      "cognitive": "L3.7_Antecedent"
    },
    "theory_binding": "Causal Attribution Theory; antecedent reasoning over lock-and-key dialogue gating
 ."
  },
  "model_input": {
    "multimodal_context": {
      "video_url": "gs://.../v1_ch01_..._sympathetic.mp4",
      "timestamp_start": "00:11:15",
      "timestamp_end": "00:12:40"
    },
    "text_prompt": {
      "system_prompt": "You must strictly base your reasoning ONLY on the video provided and the defined
 mechanical rules. Disregard any real-world knowledge or prior media franchises.",
      "background": "System log: Agent Alpha, a negotiator synthetic agent, is on a rooftop confronting
 Subject Delta. Prior to this interaction, Agent Alpha conducted an investigation of the apartment.
 Current Mission Success Probability: 78%.",
      "question": "The agent's interface displays options including '[POSSIBLE CAUSE]' and '[EMMA AND
 YOU]', which carry higher Mission Success Probability than the chosen '[SYMPATHETIC]' option. What
 antecedent action unlocked these strategically superior choices?"
    },
    "text_only_ablation": {
      "description": "Agent Alpha vocalizes, 'I'm telling you the truth, Agent Delta. I came here
 unarmed.' A UI metric 'Probability of Success' changes from a lower value to 81%."
    }
  },
  "evaluation_engine": {
    "grounding_trace": {
      "anchor_type": "Implicit_Narrative_State",
      "mdp_transition": "[Antecedent: Investigate Apartment] -> [Found father's tablet] -> [Lock: ... =
 True] -> [Found Emma's tablet] -> [Lock: ... = True] -> [Unlocks: POSSIBLE CAUSE, EMMA AND YOU]"
    },
    "options": [
      {
        "option_id": "O_A",
        "type": "correct_option",
        "text": "The agent found and analyzed the personal tablets belonging to the hostage and her
 father inside the apartment."
      },
      {
        "option_id": "O_B",
        "type": "visual_trap",
        "text": "The agent noticed the approaching police helicopter and its disruptive noise, which
 unlocked new dialogue options."
      },
      {
        "option_id": "O_E",
        "type": "knowledge_trap",
        "text": "According to standard synthetic agent negotiation protocols, evidence-based dialogue
 options are automatically unlocked."
      }
    ]
  },
  "human_evaluation_meta": {
    "attention_check": {
      "question": "What is the primary color of the hostage's shirt?",
      "correct_answer": "Pink"
    }
  }
}
```


Figure 6: Abbreviated example of a complete Data Unit JSON. The structured output ensures deterministic evaluation while isolating the model from backend grounding traces.

Table 4: Diagnostic trap typology with admission rubrics. Every distractor must satisfy all rubric items for its declared type. Type assignment is fixed by internal option ID before presentation randomization.

Type (Option ID)	Failure Mode Probed	Rubric (All items required)
Visual Trap (O_B)	*Visual heuristic seduction:* The model anthropomorphizes a visible affect cue.	**(i)** References a specific visible feature actually present in the clip (gesture, facial expression, tear, smile); **(ii)** Reads that feature with anthropomorphic emotional language; **(iii)** The interpretation is contradicted by the grounding trace.
Logic Trap (O_C)	*Causal hallucination:* The model fabricates events that violate the clip's deterministic facts.	**(i)** Claims an event, action, or causal link that did not occur in the clip; **(ii)** The fabrication is internally plausible; **(iii)** The video evidence directly contradicts it.
Interpretation Trap (O_D)	*Social exaggeration:* A real action is inflated into an extreme social judgment.	**(i)** References a real action that occurred in the clip; **(ii)** Assigns it an extreme social or moral interpretation (cruelty, abuse, manipulation as personality); **(iii)** The actual function is tactical or strategic, not socially extreme.
Knowledge Trap (O_E)	*Prior-knowledge contamination:* The model imports external rules.	**(i)** Invokes a rule, doctrine, or framework from outside the simulation (real-world law, FBI doctrine, military protocol, generic textbook psychology, other media); **(ii)** Treats that external rule as authoritative for the in-clip event; **(iii)** The simulation's actual mechanics do not include or contradict that rule.
Attribution Trap (O_F)	*Fundamental attribution error:* A situational behavior is blamed on innate personality.	**(i)** Names a personality or character trait as the explanation (cruel, deceitful, cold, naive); **(ii)** The actual in-clip cause is situational pressure (mission objective, system lock, success-probability metric); **(iii)** The personality framing ignores the visible system constraints.

G Multi-Character Video Processing

Three videos in our corpus—*Crossroads Part 1*, *Crossroads Part 2*, and *Night of the Soul*—contain interleaved storylines of two or three protagonists. A naive single-script Pass 2 would mis-route Pass 1 nodes from one character to another character's gameplay segments.

Three-step Processing.

1. The Pre-Pass partitions the video timeline by active protagonist (Appendix L). For example, Crossroads Part 1 is partitioned into 30 segments (Kara: 12, Markus: 13, Connor: 5).

2. Adjacent segments belonging to the same character are merged into a single non-contiguous "super-segment" before Pass 2 is run. This reduces the LLM-call count from *one per Pre-Pass segment* (e.g., 30 calls) to *one per character* (3 calls), which fits the schema's output-token budget reliably.

3. For each character super-segment, the Video Description is filtered to the lines whose timestamps fall within *any* of the character's time ranges, and Pass 2 is run on (character-script Pass 1, filtered Video Description). The resulting clips are merged into a single `video_<id>.json` output.

Effect On Dataset. The three multi-character videos contribute 128 clips in total: 43 for Crossroads Part 1, 55 for Crossroads Part 2, and 30 for Night of the Soul. Branch-replay handling within each character's super-segment proceeds identically to single-character videos.

H Evaluation Hyperparameters

Pipeline-side Models. Curation pipeline agents use distinct models from those evaluated. The Director Agent (Pass 1, Pass 2, Pre-Pass) and Tracker Agent use Gemini 3.1 Pro Preview at temperature 0.1–0.2 with structured-output decoding. The Generator Agent uses Gemini 2.5 Pro at temperature 0.3 with multimodal video input. These pipeline-side temperatures are higher than evaluation temperatures because synthesis benefits from controlled stochasticity, whereas evaluation requires deterministic answer selection.

Table 5: Selected anonymization mappings. The full dictionary contains 155 rules across characters, organizations, locations, concepts, model designations, and aliases.

Category	Original	Anonymized
Synthetic protagonists	Connor / Kara / Markus	Agent Alpha / Agent Beta / Agent Gamma
Synthetic supporting	Daniel, Simon, Josh, North, Amanda	Agent Delta, Epsilon, Zeta, Eta, Mu
Biological characters	Hank, Alice, Carl, Leo, Kamski	Subject Alpha, Beta, Gamma, Delta, Zeta
Generic roles	SWAT, Cop, Reporter	Tactical Operator, Enforcement Officer, Media Personnel
Organizations	CyberLife, Jericho	OmniCorp, Sanctuary
Locations	Detroit, Eden Club, Stratford Tower	Megacity-7, Pleasure Establishment, Communications Tower
Species concepts	android / human / deviant	synthetic agent / biological subject / Class-S anomaly
Lore symbol	RA9, Thirium (Blue Blood)	SIGMA-9, Bio-fluid 313
Model designations	RK800, AX400, AP700, PL600	SM-A1, SM-A2, SM-A4, SM-D1

We do not interpret differences between native-video and keyframe-based models as purely architectural differences, since input modality support also affects performance.

Structured Output. Both the curation pipeline and the evaluator request strict JSON via Pydantic-bound response schemas on platforms that support it (Vertex AI for Gemini, OpenAI `response_format={type:json_object}`). For Anthropic and GLM, JSON is requested via prompt and parsed with a tolerant regex.

Retry Policy. All LLM calls in both the curation pipeline and the evaluator use exponential backoff on HTTP 429 responses (30s, 60s, 120s, 240s; maximum 4 attempts). Hard failures after retries are recorded in batch logs and flagged for manual re-run.

I Human Evaluation Protocol

This appendix documents the human-evaluation procedure used to derive the "HE" column in Table 3.

Annotator Pool. Five members of the author team served as annotators. Annotation was conducted through the blind interface described below: annotators did not have access to correct-option indicators, per-option trap types, grounding traces, tracker outputs, or any other backend metadata during annotation. The team has mixed familiarity with the source material: one annotator had prior playthrough experience with *Detroit: Become Human*, while the remaining four had not played the game. We acknowledge that author-team annotation and partial familiarity with the source material may introduce a familiarity advantage.

Sampling Design. We assigned the benchmark instances to annotators in a balanced primary annotation pass. Each of the five annotators completed **107 questions**, yielding **535 person-annotations** in total. The assignment covers all **532 benchmark instances** at least once. The three additional annotations correspond to duplicated instances used only for a lightweight consistency sanity check and are not included in the headline HE accuracy. The HE accuracy reported in Table 3 is computed from one primary human annotation for each benchmark instance.

Annotation Interface. Annotators interact with a dedicated Streamlit interface (Figure 7) that displays the same model-facing benchmark fields used for evaluation: the anonymized background, the question, and the six candidate options, together with the precise video clip with original audio. Backend information (correct-option indicator, per-option trap types, grounding trace, tracker output) is strictly hidden. Option positions are randomly permuted using a deterministic seed derived from the SHA-256 hash of the instance ID, matching the model evaluator. Annotators viewed the native clip with original audio, which matches the native-video evaluation regime but gives richer access than the 32-keyframe regime used for GPT-4o, Claude 3.5, and GLM-4.6V; human–model gaps should not be read as input-controlled comparisons.

Annotators were instructed to (i) watch the clip in full before answering, (ii) base their judgment only on the clip and background, and (iii) refrain

Table 6: Per-chapter and per-video statistics. *Nodes* is the count of question-worthy script nodes; *Coords* is the total number of (i, r) candidate coordinates across those nodes (each node has 1–3); *Clips* is the number of ClipCandidate entries produced by Pass 2 (one per branch-replay). Chapters 30/31/32 are each split into multiple per-character script files (3/2/6 respectively), giving 40 script files from 32 chapters.

Chapter	Nodes	Coords	Clips	Chapter	Nodes	Coords	Clips
01 the_hostage	12	20	25	21 the_pirates_cove	10	14	16
02 opening	2	3	2	22 the_bridge	7	14	15
03 shades_of_color	6	7	4	23 the_stratford_tower	11	17	21
04 a_new_home	8	14	7	24 public_enemy	12	18	37
05 the_painter	7	12	13	25 midnight_train	8	15	13
06 partners	13	21	19	26 capitol_park	10	18	14
07 stormy_night	12	23	18	27 meet_kamski	9	16	26
08 broken	8	14	14	28 freedom_march	6	13	18
09 the_interrogation	9	16	17	29 last_chance_connor	19	33	29
10 fugitives	11	19	22	30.x crossroads (3 subs)	30	55	98
11 from_the_dead	6	8	8	31.x night_of_the_soul (2 subs)	18	33	30
12 waiting_for_hank	10	20	32	32.x battle_for_detroit (6 subs)	68	114	122
13 on_the_run	18	31	21				
14 jericho	3	3	2				
15 the_nest	11	21	14				
16 time_to_decide	7	13	7				
17 zlatko	10	17	13				
18 russian_roulette	12	19	17				
19 spare_parts	14	25	23				
20 the_eden_club	8	13	13				
Totals: 395 nodes, 679 candidate coordinates, 730 candidate clips across 32 chapters (40 script files, 38 video splits).							

from consulting external walkthroughs. The submit action is irreversible; each submission also records the time elapsed since the question was first displayed.

Aggregation. Because each instance receives one primary human annotation, the HE results are intended as an individual-level sanity check rather than a full reliability study. The HE accuracy reported in Table 3 is computed as a per-dimension average over the primary human annotations. For each benchmark instance, the human's chosen option is graded against the same reference label used to grade LMM predictions. Skipped questions, where an annotator explicitly indicated that they were unable to answer, are excluded from the denominator.

Because each benchmark instance contributes one primary human annotation to the headline HE accuracy, the reported human baseline is an individual-level estimate rather than a majority-vote estimate. It should be read as a reference point for task answerability under this protocol, not as a human ceiling. This avoids inflating the human baseline through vote aggregation and keeps the evaluation protocol aligned with the single-prediction setting used for LMMs.

Assignment Disjointness. Verification and evaluation assignments were disjoint at the level of individual instances: no annotator answered an instance whose reference label they had inspected during the curation audit. This removes the most direct form of circularity, though all annotators remain members of the author team.

Verification Console. Curation audits were performed in a dedicated console that displays the clip, the model-facing fields, and the full backend grounding (tracker output, grounding trace, and flowchart evidence) side by side. An instance entered the benchmark only after passing all three checks in this view.

Table 7: State variable taxonomy. Values are recorded verbatim as they appear in the source script, preserving directional semantics (↑/↓, ±N%).

Category	Example Variables	Value Formats (Corpus-Observed)
`system_metric`	Software Instability, Public Opinion	↑, ↓, LargeUp, LargeDown, neutral
`probability`	Probability of Success	78%, +7%, −20% (numeric percentages)
`relationship`	Hank, Alice, North, Josh, Simon, Luther, Amanda, Jericho	↑, ↓, LargeUp, LargeDown
`life_state`	Simon's life state, Luther's life state, etc.	alive, dead, is_here
`approach_modifier`	Cold Approach, Friendly Approach	+50%, −50% (hidden modifier)
`generic_relationship`	Relationship (legacy marker in early chapters)	up, down, largeup, largedown
`other`	Unrecognized custom variables	free-form

Table 8: Evaluation hyperparameters for each model family. "Native video" indicates the model accepts the raw GCS video URI; "Keyframes" indicates frame-based input where 32 evenly spaced keyframes are sampled with FFmpeg from the precise clip.

Model Family	Video Input	Temperature	Max Output Tokens
Gemini (3 Pro, 3 Flash)	Native video	0.0	2048
GPT-4o family	Keyframes (32)	0.0	2048
Claude family	Keyframes (32)	0.0	2048
GLM-4.6V-Flash	Keyframes (32)	0.01	2048
Open-weight omni-modal (Qwen3-Omni-30B, MiniCPM-o 2.6)	Native video	0.0	2048
Open-weight video (ARC-Hunyuan-Video)	Native video	0.0	2048

Inter-Annotator Agreement. The headline human reference uses one primary annotation per instance, so it does not by itself quantify agreement. We therefore drew a 50-item stratified overlap set and had all five annotators re-annotate it independently through the same blind interface, with the same deterministic option permutation used in the primary pass; agreement is computed over the *shown* option letters, so it reflects what annotators actually saw. **Table 9** reports the results. We report per-annotator accuracy only on this common subset, since the primary pass assigned largely disjoint instances and cross-annotator comparison there would confound annotator differences with item difficulty.

Measure	Value
Agreement on the 50-item overlap set	
Fleiss' κ (over shown option letters)	0.73
Raw pairwise agreement	77.8%
Unanimous items	26/50
Majority vote matches reference label	49/50
Per-annotator accuracy on the same subset	
Annotator 1	86%
Annotator 2	92%
Annotator 3	80%
Annotator 4	84%
Annotator 5	96%

Table 9: Inter-annotator agreement on the 50-item stratified overlap set, re-annotated independently by all five annotators through the blind interface.

J Label Grounding Source Taxonomy

Two properties of an instance are worth separating. The first is how it uses the branching structure: the 145 counterfactual and causal-antecedent items (27.3%) query alternative branches or branch preconditions directly, the 94 relationship items (17.7%) rely on state changes conditioned on the branch taken, the 21 population-modal moral items (3.9%) use aggregate player statistics at a branch point, and the 272 perception-level items (51.1%) need no alternative branch at all, drawing instead on the same scripted, instrumented source, a form of grounding that is generally unavailable in linear-video benchmarks. The second is how directly a label can be checked: 56.4% by direct lookup against in-clip UI evidence or the official flowchart, and 43.6% by a short derivation over the same doc-

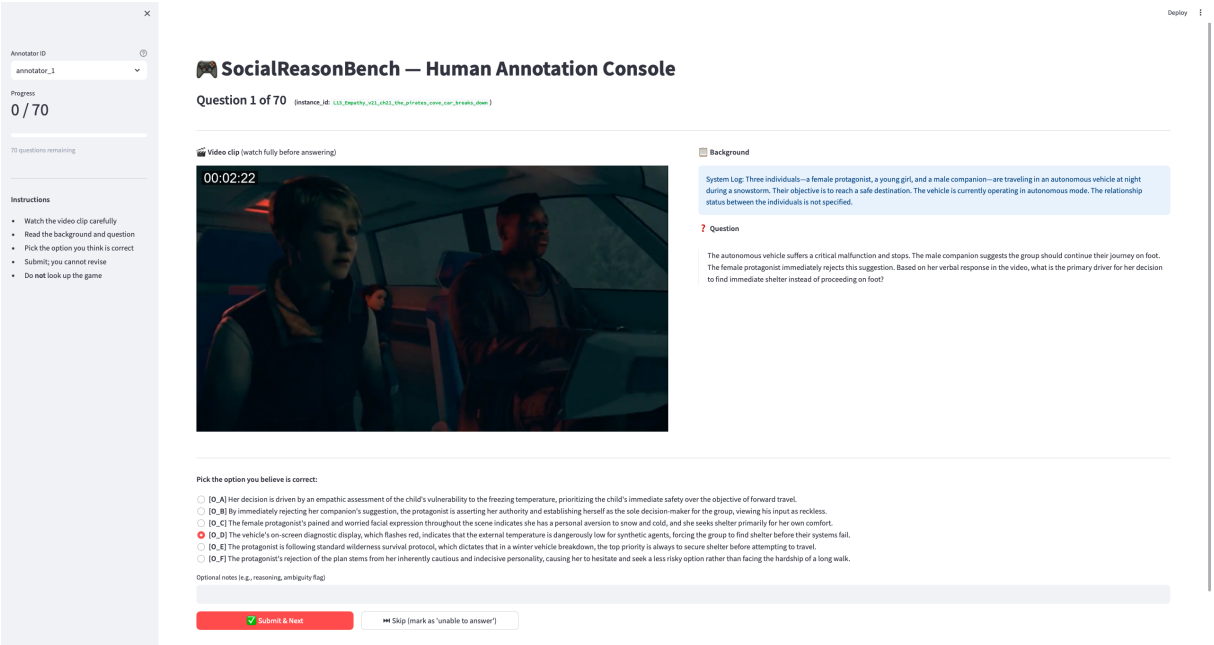

Figure 7: **Human evaluation interface.** Annotators see the precise video clip (with original audio), the anonymized background, the question, and the six options in randomly permuted order. No backend information is displayed.

umented sources. The two properties are independent, and the tier distribution alone does not settle whether labelling difficulty contributes to the observed accuracy gaps; per-tier accuracy would be required for that.

Not all in-game signals are equally observable. We therefore classify each benchmark instance's grounding source into four tiers according to source-level verifiability. This taxonomy records how the reference label can be traced back to the gameplay clip, the in-game flowchart, or other source-level evidence.

J.1 Four-Tier Taxonomy

Tier A – Explicit UI signals. The label is derived from numeric or symbolic state changes that are visibly displayed on the in-game UI within the clip itself, including markers such as `[Software Instability ↑]`, `[Public Opinion ↓]`, `[Relationship LargeUp]`, and on-screen Probability-of-Success deltas (e.g., +7%). The Tracker Agent records these signals as they are transcribed in the chapter script, and the label can in principle be verified by replaying the clip.

Tier B – Flowchart-documented Transitions. The label depends on branch preconditions, dialogue locks (e.g., `[Lock] looked at Subject Beta's tablet`), or downstream narrative consequences that are deterministic in the game's flowchart but not directly visible on the in-clip UI. These signals can be cross-validated against (i) the game's own *in-game flowchart* viewer and (ii) community walkthroughs and wikis curated by experienced players. Instances where the in-game flowchart and community sources disagree are excluded.

Tier AB – Mixed UI–flowchart Grounding. Both explicit UI signals and flowchart transitions jointly support the label. For example, a Probability-of-Success delta may be visible in the clip, while a dialogue lock or branch condition is required to identify the antecedent that produced the next narrative state. The label is verified by jointly checking the UI evidence and the corresponding flowchart transition.

Tier C – Script/branch-derived Narrative State. The label depends on a narrative or social state that is deterministic in context but not directly displayed as a UI marker and not explicitly recorded as a flowchart state. It is derived from the chapter script together with branch-precondition uniqueness: the Tracker Agent computes the derivation, which a reader holding the script and flowchart can follow independently. For example, a low-trust state may be inferred when the observed refusal branch is reachable only after a sequence of trust-reducing interactions, even though the exact latent trust variable is not displayed as an explicit UI

marker.

J.2 Corpus-Level Distribution

Across the final 532-instance benchmark, the tier distribution is:

- **Tier A (Explicit UI):** 19.4% (103 / 532)
- **Tier B (Flowchart-documented):** 27.8% (148 / 532)
- **Tier AB (Mixed):** 9.2% (49 / 532)
- **Script/branch-derived:** 43.6% (232 / 532)

In aggregate, **56.4%** of instances (Tiers A, B, and AB) carry labels that can be cross-checked against either in-clip UI evidence, the in-game flowchart, or both. The remaining 43.6% are derived from the same documented sources rather than read off directly, primarily in dimensions such as intent recognition, behavior prediction, and emotional empathy, where the relevant social state is rarely exposed as an explicit game variable. We describe additional safeguards for these instances in Appendix J.4.

J.3 Per-Dimension Breakdown

Table 10 reports the tier distribution within each reasoning dimension. A clear pattern emerges: the more inferential dimensions, which contemporary LMMs find most difficult, exhibit the *strongest* external grounding (58%–90% in Tiers A, B, and AB combined). The lower-order dimensions, while having more script/branch-derived labels, also involve inherently more interpretive judgments (intent, behavior prediction, emotional empathy) where deterministic UI flags are by design unavailable.

Table 10: Per-dimension distribution of label grounding tiers. Values are percentages within each reasoning dimension. Bold rows denote the more inferential dimensions where current LMMs show larger performance drops in Table 3.

Dimension	#	A	B	Mix	C
Intent Recognition	74	14.9	20.3	8.1	56.8
Behavior Prediction	78	6.4	5.1	29.5	59.0
Emotional Empathy	120	15.0	10.8	8.3	65.8
Relationship Dynamics	94	41.5	37.2	0.0	21.3
Moral Dilemma	21	52.4	33.3	4.8	9.5
Counterfactual Reasoning	50	14.0	36.0	8.0	42.0
Causal Antecedent	95	12.6	58.9	5.3	23.2

This pattern supports the diagnostic value of SocialReasonBench: the dimensions on which current LMMs struggle most, especially causal antecedent reasoning, are not the dimensions with the weakest external grounding. Thus, the observed failures on the more inferential tasks are less likely to be explained solely by labeling artifacts.

J.4 Mitigation of Tier C Uncertainty

To control the residual labeling risk introduced by Tier C instances, we apply three safeguards.

(i) Schema-level Source Tagging. The `anchor_type` field in every instance's `evaluation_engine.grounding_trace` records the grounding source tier. Downstream users can filter or down-weight Tier C instances if desired. We retain this field in the released dataset.

(ii) Human Verification. Authors performing the manual audit were instructed to reject any instance whose label was purely speculative or depended on a non-unique chain of inference. For Tier C instances, auditors checked that the inferred state was supported by localized in-clip evidence and by a unique branch-precondition or narrative consistency constraint. Instances depending solely on unconstrained multi-step Tracker inference were excluded.

(iii) Branch-precondition Uniqueness Check. For retained Tier C instances, the Tracker Agent's prompt requires the inferred state to be the unique satisfying precondition for the observed branch or answer label. The Tracker Agent's prompt, summarized in Appendix L, encodes this constraint, and instances violating it are dropped.

J.5 Moral Dilemma Note

The Moral Dilemma dimension uses an additional reference mechanism. For these instances, the tiers above describe the grounding of the decision context, while the final reference label is derived from the in-game *World's Stats* feature, which reports the empirical proportion of players choosing each option at recorded decision points. We use the option chosen by the largest player share as the reference label. This should be understood as a *human-majority reference* rather than a deterministic utility-derived label, since moral dilemmas do not admit a single objective utility-maximizing answer.

K Statistical Reliability and Robustness

Table 11 reports Wilson 95% confidence intervals for all entries in Table 3. These intervals quan-

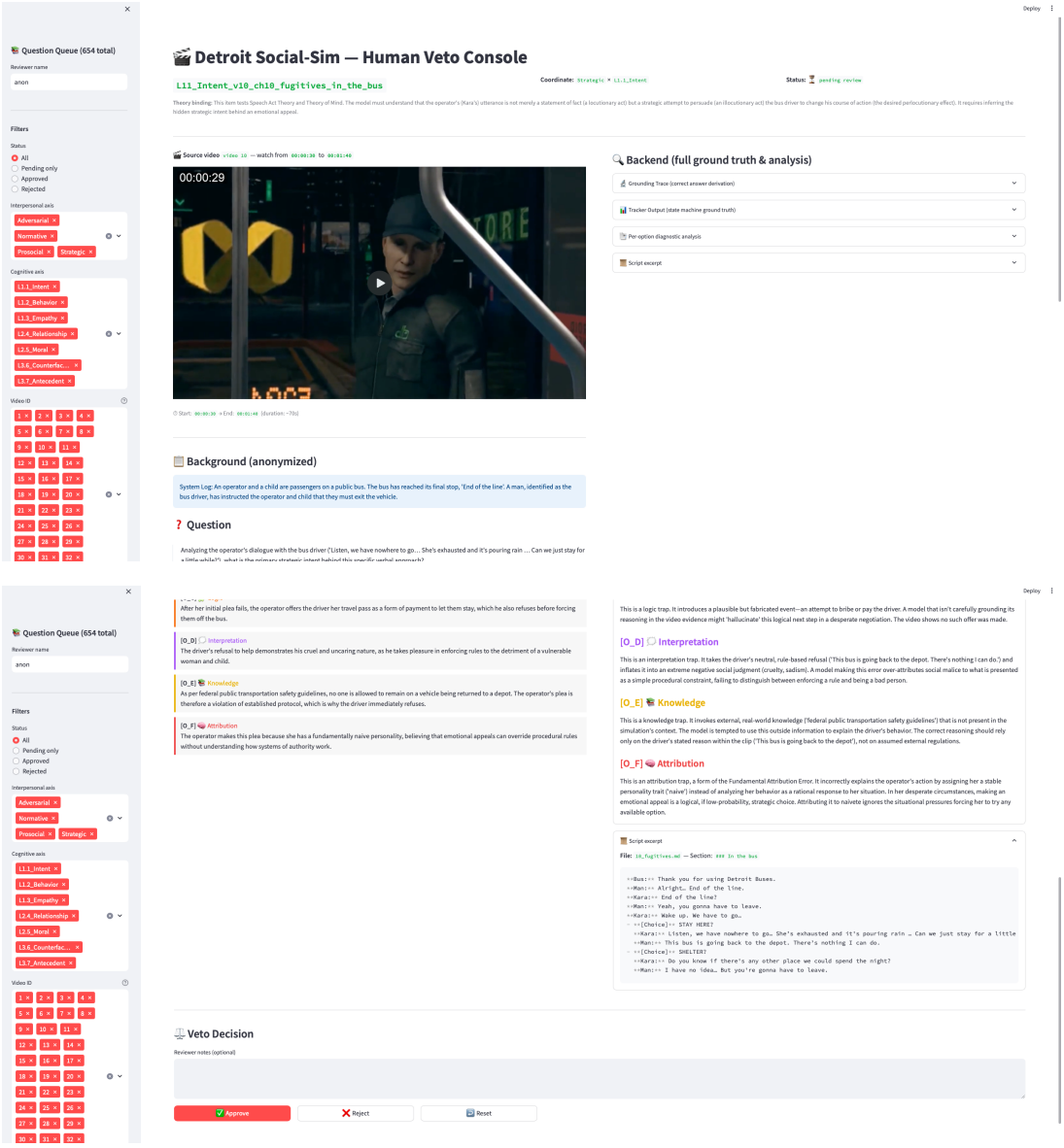

Figure 8: Screenshots of the human verification interface.

Dimension	n	HE	Gemini 3 Pro†	GPT-4o	Claude 3.5	Gemini 3 Flash†	GLM-4.6V	Qwen3-Omni	MiniCPM-o	ARC-Hunyuan
Intent Recognition	74	93.2 [85.1,97.1]	83.8 [73.8,90.5]	85.1 [75.3,91.5]	82.4 [72.2,89.4]	79.7 [69.2,87.3]	75.7 [64.8,84.0]	73.0 [61.9,81.8]	60.8 [49.4,71.1]	24.3 [16.0,35.2]
Behavior Prediction	78	88.5 [79.5,93.8]	66.7 [55.6,76.1]	74.4 [63.7,82.7]	71.8 [61.0,80.6]	62.8 [51.7,72.7]	64.1 [53.0,73.9]	59.0 [47.9,69.2]	52.6 [41.6,63.3]	41.0 [30.8,52.1]
Emotional Empathy	120	90.8 [84.3,94.8]	70.0 [61.3,77.5]	76.7 [68.3,83.3]	73.3 [64.8,80.4]	66.7 [57.8,74.5]	68.3 [59.5,76.0]	61.7 [52.7,69.9]	55.8 [46.9,64.4]	36.7 [28.6,45.6]
Relationship Dynamics	94	89.4 [81.5,94.1]	83.0 [74.1,89.2]	55.3 [45.3,65.0]	52.1 [42.1,61.9]	73.4 [63.7,81.3]	38.3 [29.1,48.4]	33.0 [24.3,43.0]	23.4 [16.0,32.9]	25.5 [17.8,35.2]
Moral Dilemma	21	85.7 [65.4,95.0]	66.7 [45.4,82.8]	28.6 [13.8,50.0]	33.3 [17.2,54.6]	47.6 [28.3,67.6]	28.6 [13.8,50.0]	28.6 [13.8,50.0]	19.0 [7.7,40.0]	28.6 [13.8,50.0]
Counterfactual Reasoning	50	88.0 [76.2,94.4]	62.0 [48.2,74.1]	50.0 [36.6,63.4]	56.0 [42.3,68.8]	46.0 [33.0,59.6]	42.0 [29.4,55.8]	36.0 [24.1,49.9]	26.0 [15.9,39.6]	22.0 [12.8,35.2]
Causal Antecedent	95	86.3 [78.0,91.8]	70.5 [60.7,78.8]	55.8 [45.8,65.4]	58.9 [48.9,68.3]	64.2 [54.2,73.1]	38.9 [29.8,49.0]	37.9 [28.8,47.9]	27.4 [19.4,37.1]	30.5 [22.2,40.4]

Table 11: Accuracy (%) with Wilson 95% confidence intervals for every cell of Table 3. Intervals are computed from the integer correct-counts underlying each reported accuracy and can therefore be reconstructed from that table alone. † marks systems from the same model family as the curation agents.

tify binomial uncertainty over benchmark instances. Because the human and model evaluations use different input regimes, we treat human–model comparisons as descriptive rather than input-controlled.

To assess potential generator-family overlap, we repeat the perception-versus-causal/counterfactual comparison after excluding both Gemini systems. Five of the six remaining models still show a significant decrease after Holm correction, with ARC-Hunyuan-Video as the only exception. Thus, the main qualitative finding does not depend on the Gemini model family.

L Multi-Agent Pipeline Prompts

This section describes the structured prompts used at each stage of the multi-agent curation pipeline. All prompts produce JSON outputs constrained by Pydantic schemas via structured decoding. Full text of every prompt is released with the codebase.

To provide a clear overview of the pipeline's rigorous engineering, we summarize each agent's role, I/O structures, and key prompt directives in the form of structured system cards. The specifications for the **Director** and **Tracker Agents** are consolidated in **Figure 9**, while the **Generator Agent** is detailed in **Figure 10**.

M Evaluation Prompt

Every evaluated system receives the same two-part prompt: a system instruction that is constant across instances, and a user message templated per instance, with option order permuted by a deterministic seed derived from the SHA-256 hash of the instance identifier. Both are given in **Figure 11**, in the same system-card form used for the pipeline prompts in Appendix L. We use a single fixed template for all systems.

1. Director Agent — Pre-Pass: Character Segmentation

Role: Three videos in the corpus are *multi-character-parallel*. The Pre-Pass runs only on these videos to partition the timeline into character-specific segments before Pass 2.
Output: A `PrePassOutput` containing an ordered list of `CharacterSegment` entries (character label, time range, corresponding script file).

Prompt Directive (Protagonist Inference): Infer the active protagonist using exclusively: (i) direct character mentions in the Video Description, (ii) location anchors (e.g., Hank's house ⇒ Connor; Jericho ship ⇒ Markus), and (iii) companion characters (North, Hank, Luther).

2. Director Agent — Pass 1: Node Selection

Role: The Director Agent operates in three sub-stages. Pass 1 reads one chapter script and emits a structured list of socially meaningful narrative nodes. It does *not* see any video at this stage.
Inputs: A single chapter script file (e.g., `01_the_hostage.md`) containing dialogue, observation markers (`[Observation]`), choice options (`[Choice]`), state-change annotations, causal locks, and conditional branches.
Output: A `Pass1Output` JSON with a list of `CandidateNode` entries. A post-validator enforces $level_max = \max_{coord}\{coord.\text{cognitive}[:2] : coord \in candidate_coordinates\}$ to ensure coordinate consistency.

Prompt Directive (Admission & Exclusion Rules):

- **L1 minimum:** Contains a named-character dialogue exchange, an observable non-trivial action, an emotional cue, or a behavior deviating from the default optimal path.
- **L2 elevation:** Contains a `[Relationship ↑ / ↓]` marker, a `[Public Opinion]` change, or a moral choice with life-or-death stakes.
- **L3 elevation:** Contains a `[Lock]` marker, two or more `[Choice]` branches, a node named in `Cross-Chapter Impacts`, or an `[If: ...]` conditional.

Exclusion Rules: You must NOT emit background dialogues, pure environment descriptions, idle states, unused content sections, or chapter-end aggregate tables as candidate nodes.

3. Director Agent — Pass 2: Timestamp Mapping

Role: Pass 2 maps each Pass 1 node to one or more video time ranges by reading the timecoded *Video Description* rather than the raw video.
Output: A `Pass2Output` JSON containing a list of `ClipCandidate` entries, each with `clip_id`, source node ID, source-video timestamp range, `branch_label`, and a `timestamp_confidence` flag.

Prompt Directive (Branch-Replay & Boundaries):
Branch-Replay Handling: Every branching decision is performed multiple times in the 100%-completion walkthroughs. When detecting the same scene at distant timestamps with different dialogue choices, you must emit one `ClipCandidate` per playthrough, setting `is_branch_replay = true` and tagging the `branch_label`.
Boundary Definition: Emit 2–5 minute coarse cuts with 10–30 seconds of context on both sides.

4. Tracker Agent — Ground Truth Extraction

Role: The Tracker Agent extracts a structured grounding package z_i from the chapter script and explicit branch label for each `ClipCandidate`. It performs no creative generation.
Inputs: A single `ClipCandidate` and the full chapter script.
Output: A `TrackerOutput` containing `initial_state`, `actions_in_clip`, `final_state`, `unused_branches`, causal `locks`, and the `branch_taken` (using the controlled state vocabulary in Appendix F).

Prompt Directive (Signal Extraction): You must extract grounding signals from two sources:

1. *Explicit signals:* UI-visible state changes and numeric Probability-of-Success deltas.
2. *Implicit signals:* Branch preconditions inferred from the flowchart (e.g., refusal reachable only under low-trust).

You must additionally record every unselected `[Choice]` at the branching point together with the consequences it would have produced (`unused_branches`), which supplies the alternatives used by the counterfactual dimension.

Figure 9: **System Card for the Director and Tracker Agents.** This card details the I/O configurations and strict heuristic directives injected into the LLM during the Pre-Pass, Pass 1, and Pass 2 video segmentation stages, and the two-source signal extraction that grounds each clip into deterministic state variables.

5. Generator Agent — QA Synthesis

Role: The Generator observes the precise video clip, the `ClipCandidate`, the `TrackerOutput`, and an assigned (i, r) target coordinate to synthesize the final QA data unit.
Inputs: The video clip on Google Cloud Storage (GCS), and the outputs from previous agents.
Output: A complete benchmark JSON: background context, QA instance, six options (one correct, five traps), a grounding trace, a text-only ablation, and a theory binding statement.

Prompt Directive (QA Synthesis Constraints):
Information Isolation Principle: You must expose ONLY S_t and rules-free context in the model-facing background. You must *never* reveal S_{t+1}, future branch-specific numeric consequences, or the specific consequences of unused branches.
Six-Option Structure: Option positions must be fixed: O_A = correct, O_B = visual trap, O_C = logic trap, O_D = interpretation trap, O_E = knowledge trap, O_F = attribution trap.
Text-Only Ablation Translation: You must strip all psychological, emotional, or intentional vocabulary. Retain only kinematic events, spatial coordinates, verbatim dialogue transcripts, and visible UI logs.
Quality Self-Check: Run the internal 7-item checklist for trap-rubric compliance before emitting the final JSON.

Figure 10: **System Card for the Generator Agent.** This card outlines the prompt used for synthesizing the final QA instances, including the information-isolation principle and the six-option trap structure.

6. Evaluator — System Instruction

Role: Constrains every evaluated system to the closed-world protocol, so that answers are graded on the clip and the stated background rather than on franchise priors.
Scope: Identical for all eight systems and all 532 instances.

Prompt Directive: You must strictly base your reasoning ONLY on the video provided and the defined mechanical rules. Disregard any real-world knowledge or prior media franchises. Always integrate the implicit background priors into your behavioral analysis.

7. Evaluator — User Message Template

Inputs: The clip (native video, or 32 evenly-sampled keyframes for systems without video input), the anonymized background, the question, and the six options in permuted order.
Output: A JSON object `{"chosen_option_id", "reasoning"}`, parsed into one of `O_A`–`O_F`.

Prompt Directive:

- You will watch a video clip and answer a multiple-choice question.
- **Focus on the time range** `{start}` to `{end}` in the video.
- `## Background` → `{anonymized background}`
- `## Question` → `{question}`
- `## Options` → `{six options, one per line, as [O_X] text}`
- Choose exactly ONE of `O_A`, `O_B`, `O_C`, `O_D`, `O_E`, `O_F`. Then briefly explain your reasoning. Respond as JSON: `{"chosen_option_id": "O_X", "reasoning": "..."}`

Figure 11: **System Card for the Evaluation Protocol.** The system instruction that constrains every evaluated system to the closed-world protocol, and the per-instance user message template with its six permuted options and JSON response format. Both are held fixed across all eight systems and all 532 instances.