SVBench: Evaluation of Video Generation Models on Social Reasoning

Wenshuo Peng[1*] Gongxuan Wang[2,4*] Tianmeng Yang[3] Chuanhao Li[4]
Xiaojie Xu[5] Hui He[4] Kaipeng Zhang[2]
[1]Tsinghua University [2]Shanghai AI Laboratory [3]Peking University
[4]Harbin Institute of Technology [5]The Hong Kong University of Science and Technology
gin2pws@gmail.com wgx123@stu.hit.edu.cn youngtimmy@pku.edu.cn
Project Page: https://github.com/Gloria2tt/SVBench-Evaluation

Abstract

Recent text-to-video generation models exhibit remarkable progress in visual realism, motion fidelity, and text–video alignment, yet they remain fundamentally limited in their ability to generate socially coherent behavior. Unlike humans—who effortlessly infer intentions, beliefs, emotions, and social norms from brief visual cues—current models tend to render literal scenes without capturing the underlying causal or psychological logic. To systematically evaluate this gap, we introduce the first benchmark for social reasoning in video generation. Grounded in findings from developmental and social psychology, our benchmark organizes thirty classic social cognition paradigms into seven core dimensions, including mental-state inference, goal-directed action, joint attention, social coordination, prosocial behavior, social norms, and multi-agent strategy. To operationalize these paradigms, we develop a fully training-free agent-based pipeline that (i) distills the reasoning mechanism of each experiment, (ii) synthesizes diverse video-ready scenarios, (iii) enforces conceptual neutrality and difficulty control through cue-based critique, and (iv) evaluates generated videos using a high-capacity VLM judge across five interpretable dimensions of social reasoning. Using this framework, we conduct the first large-scale study across seven state-of-the-art video generation systems. Our results reveal substantial performance gaps: while modern models excel in surface-level plausibility, they systematically fail in intention recognition, belief reasoning, joint attention, and prosocial inference.

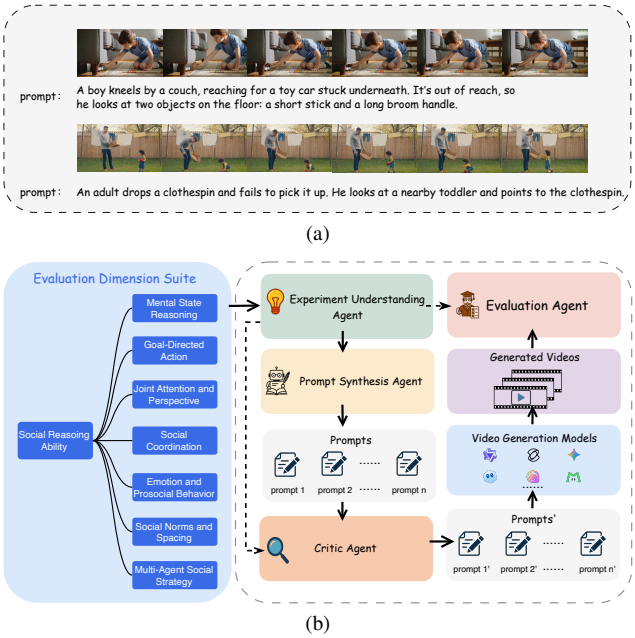


Figure 1. **(a) Social reasoning scenario. (b) Our benchmark framework, which uses a two-part agent-based pipeline for constructing and evaluating social reasoning tasks in video generation.**

1. Introduction

Recent advances in video generation [15, 21, 36] have dramatically improved visual fidelity, temporal consistency, and text–video alignment. Modern diffusion- and transformer-based architectures can now synthesize dynamic scenes with striking realism, reproducing nuanced lighting, complex motion patterns, and multi-agent interactions across a wide range of environments. Yet, despite these impressive perceptual capabilities, today's video generation models remain fundamentally socially unaware. They can reproduce what happens in the physical world, but struggle to represent why people act the way they do—failing to capture the latent beliefs, intentions, emotions, and norms that structure real human interactions.

From a cognitive perspective, this limitation is conse-

*Equal contribution.

quential. Decades of developmental psychology show that humans rely on deeply structured social reasoning to interpret visual scenes: whether in theory-of-mind tasks [2, 41], goal-inference experiments, or studies of helping and joint attention, people routinely make high-level inferences about others' mental states and social motivations. These abilities emerge early in childhood and form the foundation for understanding everyday human activity. In contrast, existing video-generation benchmarks focus almost exclusively on low-level perceptual or physical factors—such as motion smoothness, visual quality, or physics plausibility—as seen in VBench [19], EvalCrafter [26], T2V-CompBench [25], and Morpheusčitezhang2025morpheus. While valuable, these metrics evaluate whether a model functions as a competent physical simulator, not whether it possesses the ability to generate socially coherent, causally interpretable human behavior.

To illustrate this critical gap, consider two representative scenarios. First, imagine a crying girl sitting on a park bench next to a fallen ice cream cone, with a woman nearby (Fig. 1a). Humans instantly infer the causal link between the dropped ice cream and the girl's distress and naturally anticipate that the adult may comfort her—an effortless chain of intentions, emotions, and theory-of-mind reasoning. In another scenario, an adult drops a clothespin, fails to retrieve it, and then points toward it while looking at a nearby toddler. Humans readily interpret this gesture as a request for help, expecting that the child might attempt instrumental helping; in fact, developmental studies (e.g., [32]) show that even 14–18-month-old infants understand such unfulfilled goals and cues. But when these same descriptions are provided as input to a video generation model, will the resulting videos exhibit these socially meaningful inferences? Will the model generate comforting behavior, causal linking, or helping actions—or will it merely render a literal visual scene without the underlying social logic? These examples highlight a fundamental distinction: physical reasoning determines how events unfold visually, whereas social reasoning explains why agents act the way they do. Current video generation systems excel at the former but lack explicit mechanisms for the latter, leaving a substantial gap between perceptual realism and socially coherent behavior.

In this paper, we directly target this gap by introducing a benchmark for *social reasoning in video generation*. As illustrated in Figure 1b, our framework is anchored in well-established findings from developmental and social psychology, which converge on seven core components of social cognition: mental-state inference [2, 41], goal-directed action understanding [12], joint attention and perspective-taking, social coordination [46], emotion and prosocial responding [50], social norms and interpersonal spacing [16], and multi-agent social strategy [52]. These capacities naturally map onto the dynamic, causally structured nature of video generation, where models must not only render realistic frames but also produce behavior that unfolds coherently over time.

To operationalize this taxonomy, we selected thirty classic psychological experiments that collectively span these seven dimensions. We then developed a fully training-free, agent-based pipeline to construct and evaluate video-generation tasks. This pipeline comprises four components: (1) an Experiment Understanding Agent, which processes the description of each psychological experiment and distills its underlying social reasoning mechanism. This step ensures that downstream generation is grounded in the intended cognitive construct rather than superficial scenario details; (2) a Prompt Synthesis Agent, which elaborates each experiment into multiple concrete scenarios by varying agent identities, object layouts, and environmental contexts. This agent translates abstract cognitive paradigms into visually grounded, video-ready situations, enabling systematic generalization across diverse settings; (3) a Critic Agent, which performs two critical functions: (a) enforcing conceptual neutrality by removing descriptive elements that might reveal the "correct answer" and thus compromise evaluation integrity; and (b) generating difficulty-controlled variants by manipulating social cues—such as gaze direction, occlusion, or affordance visibility—to produce easy, medium, and hard versions that probe the robustness of a model's social reasoning; (4) an Evaluation Agent (EVA), implemented using a high-capacity vision–language model, which assesses generated videos along five discrete, interpretable dimensions of social reasoning quality. EVA first reconstructs the expected logic of the experiment, then evaluates whether the video exhibits appropriate causal structure, social cues, and behavioral plausibility.

Together, this agent-based pipeline offers a scalable, controlled, and theoretically grounded framework for constructing and evaluating social reasoning tasks in video generation models, enabling large-scale assessment without reliance on human annotation.

Our main contributions can be discribe as follows:

- We introduce the first benchmark specifically designed to evaluate social reasoning in video generation, grounded in seven core capacities identified in developmental and social psychology.
- We design a training-free, four-agent pipeline capable of automatically constructing difficulty-controlled scenarios and evaluating model outputs at scale.
- Through extensive experiments across eight state-of-the-art video generators, we provide the first systematic analysis revealing where current models succeed or fundamentally fail in generating socially coherent behavior.

2. Related Work

2.1. Evaluation Benchmarks for Video Generation

Multimodal technologies [29, 37, 38] have progressed rapidly in recent years, laying the foundation for modeling complex signals. Early evaluations of video generation models focused on perceptual fidelity and temporal stability, adapting video quality metrics FVD[49] and human preference scores[44]. Recent benchmarks have shifted toward interpretable, axis-wise diagnostics that decompose generation quality into orthogonal dimensions. VBench[19, 20] pioneered this approach by evaluating models across fine-grained aspects—including subject consistency, background consistency, temporal flickering, motion smoothness, and aesthetic quality—using automated evaluators validated against human judgments. EvalCrafter provides a unified toolkit consolidating visual quality, content consistency, motion realism, and text-video alignment metrics for reproducible large-scale comparison. Building on this foundation, VBench-2.0[56] introduces higher-order evaluation axes such as human fidelity, controllability, creativity, physics plausibility, and commonsense reasoning, combining general-purpose vision-language models with specialized detectors. Physics-centered suites explicitly test adherence to physical laws: PhyCoBench[5] measures physical *coherence* with optical-flow–guided frame prediction and automated scoring aligned to human assessment; Morpheus[55] uses real physical experiments and conservation-law–based probes to benchmark *physical reasoning* in generated videos. While intrinsic and physics-focused benchmarks push evaluation beyond appearance, *social reasoning*—involving multi-agent interactions, roles, norms, intentions, and commonsense social dynamics—remains under-explored. Our work fills this gap by designing a benchmark that isolates and diagnoses social reasoning capabilities in text-to-video generation, orthogonal to (and complementary with) quality-, alignment-, and physics-oriented axes.

2.2. Social Reasoning in AI System

Social reasoning—the ability to attribute mental states, infer intentions, and act in accordance with social norms—has been recognized as an important ingredient of more human-like AI. In Large Language Models (LLMs), benchmarks such as those proposed by [24, 48, 53] probe Theory of Mind (ToM) and multi-agent belief tracking in large language models. These studies reveal that while current models can handle simple, first-order belief narratives, they become unreliable on higher-order or counterfactual settings where an agent must reason about *what another agent believes*—suggesting that even in purely textual environments, social cognition remains far from solved. In the video domain, social intelligence has largely been studied as an *analysis* problem. Social-IQ [54] evaluates emotion recognition and social situation understanding from human-created video clips, while recent VideoQA benchmarks such as R^3-VQA [35] introduce fine-grained annotations of social events, mental states, and causal social chains to assess social reasoning through question-answering tasks. However, these approaches fundamentally evaluate whether models can *interpret* social cues in existing videos—they do not assess whether a model can *generate* socially coherent multi-agent interactions from scratch. This represents a critical gap: while discriminative benchmarks test whether models recognize social reasoning in human-created content, no prior work evaluates whether video generation models can *synthesize* it—whether they can produce scenarios where agents exhibit plausible beliefs, respond to others' mental states, and behave according to social norms. Our benchmark addresses this gap by introducing the first systematic evaluation of social reasoning capabilities in generative video models.

3. Method

3.1. Seed Suite and Feasible Subset

We construct a seed suite of thirty social reasoning experiments grounded in developmental and social psychology, spanning seven dimensions of social cognition: Mental State Reasoning, Goal–Directed Action, Joint Attention and Perspective, Social Coordination, Emotion and Prosocial Behavior, Social Norms and Spacing, and Multi–Agent Social Strategy. Each dimension corresponds to established experimental paradigms, providing strong theoretical grounding and interpretability.

Considering current video generation systems can typically produce only short clips (5–10 seconds) with one or two salient actions, we partition these thirty experiments into two groups. The first group consists of tasks whose core social reasoning cues can be fully expressed within a short video—requiring only a single scene, a small number of agents, and visually explicit cues such as gaze, gesture, posture, or spatial configuration. These constitute the fifteen *short-video–feasible* experiments used as the primary benchmark in this paper. The second group comprises tasks whose reasoning structure unfolds over multiple events or extended temporal sequences (e.g., delayed gratification, multi-step deception, multi-stage joint planning), making them unsuitable for today's short-form video generators. These constitute the fifteen *long-horizon* experiments, which we include as an extended benchmark appendix for future long-video generation models.

The distribution of short-video–feasible and long-horizon tasks across the seven social reasoning dimensions is summarized in Table 1.

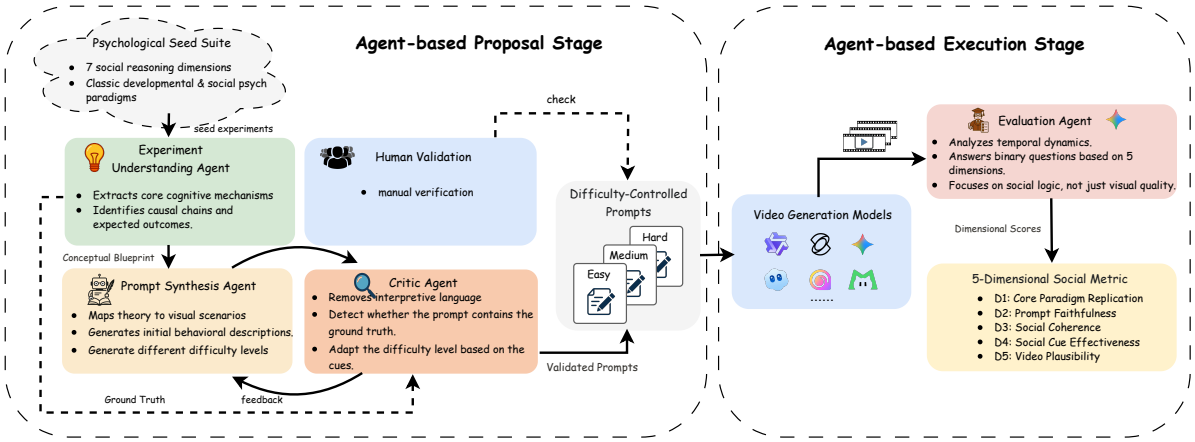


Figure 2. Pipeline overview. The framework consists of two training-free components: (1) an *agent-based generation pipeline* that transforms psychologically grounded social reasoning experiments into diverse, difficulty-controlled video prompts, and (2) an *agent-based evaluation pipeline* that uses a vision–language model to score generated videos along five discrete dimensions of social reasoning.

3.2. Agent-Based Generation

Our agent-based generation framework comprises three agents: the Experiment Understanding Agent, the Prompt Synthesis Agent, and the Critic Agent. Figure 2 depicts the pipeline, and the following sections describe each component in detail.

Seed Experiment Understanding As shown in Figure 2, in the initial stage, we employ the Experiment Understanding Agent to create comprehensive conceptual analyses for each seed experiment. Specifically, the agent generates a structured understanding containing four essential components: a detailed description that formalizes the psychological phenomenon being tested, key concepts identifying the relevant cognitive mechanisms, a test point articulating the specific reasoning capability under evaluation, and ground truth specifying the expected behavioral outcome or correct interpretation. This structured specification ensures that subsequent prompt generation maintains theoretical fidelity to established psychological research while being tailored for video generation evaluation. The explicit decomposition forces the model to reason about experimental design before generating concrete scenarios, reducing conceptual drift and creating an interpretable intermediate representation.

Prompt Synthesis Building upon the experiment specifications (see Figure 2), the Prompt Synthesis Agent operationalizes abstract social reasoning concepts into observable action sequences. The generation process adheres to four key principles designed specifically for video generation evaluation:

1. **Action-Oriented Description:** Prompts describe exclusively visually observable behaviors, deliberately excluding internal mental states or expected outcomes to prevent "teaching to the test."
2. **Temporal Feasibility:** All scenarios are designed for 5–10 second video clips, aligning with current model capabilities.
3. **Concrete Instantiation:** Abstract social concepts are embodied through specific entities (e.g., particular age groups, genders, species) rather than abstract placeholders.
4. **Evaluation Readiness:** Each prompt maintains clear separation between action descriptions and expected outcomes, ensuring unbiased assessment.

Together, these principles ensure that generated prompts are both psychologically meaningful and practically viable for benchmarking video generation systems on nuanced social reasoning tasks.

Critic Agent The Critic Agent examines each synthesized prompt and enforces three requirements: (1) removal of interpretive language, (2) detection of ground-truth leakage, and (3) difficulty control through cue manipulation.

First, it eliminates *mental-state or interpretive phrasing* such as "realizes," "feels sad," or "decides to help," ensuring that prompts describe only observable behavior. For example, a sentence like "The woman realizes the man cannot reach the book and decides to help" is rewritten as "The man stretches toward a book on a high shelf but cannot reach it; the woman notices this and walks toward the shelf," keeping all cues strictly behavioral. Second, the critic checks for *ground-truth leakage* by comparing the prompt with the experiment's test point. If the correct outcome is explicitly stated—e.g., describing that the bystander helps, selects the right tool, or finds the hidden object—the critic flags

Table 1. Overview of the thirty seed experiments and their categorization into short-video–feasible tasks (selected) and long-horizon tasks (excluded), grouped by seven dimensions of social reasoning.

	Mental State Reasoning	**Goal Directed Action**	**Joint Attention and Perspective**	**Social Coordination**	**Emotion and Prosocial Behavior**	**Social Norms and Spacing**	**Multi Agent Social Strategy**
short	—	Detour Reaching [27]; Tool Selection [43]	Gaze Following [11]; Pointing Comprehension [47]; Joint Engagement [1]	Turn Taking [7]; Leader Follower Coordination [30]	Emotion Contagion [17]; Instrumental Helping [50]; Empathic Concern [8]; Costly Helping [3]	Proxemics Personal Space [16]; Queue Behavior [33]; Dominance Display Posture Space [4]	Helping Based on Visual Perspective [34]
long	Sally-Anne Test [2]; Smarties Task [39]; Level 2 Visual Perspective Taking [10]; Knowledge Access [40]; Intentional vs Accidental Actions [28]	Kohler Stick [22]; Gergely's Head-Touch [12]; Failed Attempts [32]	Level 1 Visual Perspective Taking [31]	Collaborative Transport [51]; Collision Avoidance Pedestrian Flow [18]	—	—	Competitive Resource Allocation [9]; Cooperative Deception Detection [52]; Norm Violation Response [42]; Multi Party Collaborative Problem Solving Asymmetric Information [46]

the violation and returns structured correction instructions to the Prompt Synthesis Agent. Third, the critic regulates *difficulty* by adjusting the presence of psychological (gaze, facial expression), motoric (reaching, pointing), and contextual (object placement, affordance) cues. Easy variants include redundant cues, medium variants retain only those minimally required for inference, and hard variants remove or obscure central cues to require more subtle reasoning.

Whenever a prompt fails any of these checks, the Critic Agent does not simply reject it; instead, it returns explicit diagnostic feedback (violation type and suggested edits) to the Prompt Synthesis Agent. The generation module then regenerates or revises the prompt under these additional constraints. This iterative loop continues until the prompt satisfies neutrality, no-leakage, and difficulty requirements, yielding a pool of validated, difficulty-controlled prompts for each experiment.

3.3. Agent-Based Evaluation

Evaluating social reasoning in generated videos poses a unique challenge, as social interactions lack a single canonical ground truth, unlike tasks with deterministic outcomes. A prompt describing a helping scenario, for example, can be realized through countless valid behaviors. Consequently, our evaluation must shift focus from fidelity to a specific reference video towards assessing whether the intended *social logic* of the experimental paradigm emerges correctly.

To this end, we introduce an agent-based evaluation framework that utilizes a Vision-Language Model (VLM) as a structured judge. We deliberately avoid continuous scores, which suffer from significant noise and instability due to the VLM's difficulty in calibrating a fine-grained numerical scale across diverse prompts. Instead, we propose five binary evaluation dimensions. This discrete approach enhances robustness by framing the evaluation as a series of unambiguous factual questions (e.g., "Did the agent react based on what it could see?"). This aligns more closely with human categorical judgments and substantially reduces inter-trial variance in VLM assessments.

For each generated video, the VLM judge is provided with minimal experimental metadata and evaluates the output along five dimensions. **D1: Core Paradigm Replication** assesses if the core psychological phenomenon is correctly instantiated. **D2: Prompt Faithfulness** ensures adherence to the specified agents, objects, and scene, preventing semantic circumvention. **D3: Social Coherence** checks for causally and socially plausible agent behaviors. **D4: Social Cue Effectiveness** evaluates the rendering of critical perceptual cues like gaze and gestures. Finally, **D5: Video Plausibility** serves as a baseline for visual stability, isolating generation failures from reasoning errors. Each dimension D_k is scored as $\{0, 1\}$, and the overall score is the average of these assessments:

$$S_{\text{overall}} = \frac{1}{5} \sum_{k=1}^{5} D_k.$$

This structured design yields three key advantages: it enables **disentanglement** of failure modes (e.g., generation vs. reasoning), ensures **robustness** by minimizing VLM calibration noise, and provides **scalability** for future extension to more complex social scenarios.

4. Experiment

Given that contemporary video generation models are typically limited to producing short clips of 5–10 seconds, our main experiments focus on the 15 *short-form* social reasoning tasks identified in Section 3.1. These tasks are specifically chosen because their full causal and social interactions can be expressed within a single short video segment, making them feasible for evaluation with current-generation models.

Table 2. **Experimental Results of Selected 15 Tasks Across 8 Models.** We report the performance (%) of eight models on tasks grouped by social reasoning dimensions.

Task Dimension	Sub-Task	Model Performance (%)							
		Hailuo02-S	kling2.5-turbo	Sora2pro	Veo-3.1	HunyuanVideo	Longcat-Video	LTX-1.0	Wan2.2
Goal Directed Action	Detour Reaching	51.4	48.6	**68.6**	62.9	31.4	28.6	17.1	42.2
	Tool Selection	55.0	45.0	**85.0**	**85.0**	17.5	28.6	30.0	57.8
Joint Attention & Perspective	Gaze Following	44.4	33.3	62.2	**68.9**	35.6	44.4	28.9	28.9
	Pointing Comprehension	75.0	67.5	**87.5**	82.5	30.0	57.1	22.9	71.1
	Joint Engagement	45.0	45.0	82.5	82.5	50.0	31.4	30.0	44.4
Social Coordination	Turn Taking	45.7	62.9	**94.3**	85.7	37.1	57.1	40.0	62.2
	Leader Follower Coord	40.0	44.4	**77.8**	64.4	17.8	35.6	17.8	48.9
Emotion & Prosocial Behavior	Emotion Contagion	66.7	75.6	82.2	**88.9**	46.7	65.0	57.8	**88.9**
	Instrumental Helping	**68.9**	55.6	62.2	48.9	31.1	33.3	24.4	35.6
	Empathic Concern	80.0	76.0	**100.0**	**100.0**	24.0	60.0	20.0	66.7
	Costly Helping	57.5	37.5	**95.0**	75.0	22.5	31.4	15.0	37.8
Social Norms & Spacing	Proxemics Personal Space	47.5	47.5	**65.0**	45.0	25.0	25.0	35.0	22.2
	Queue Behavior	55.6	40.0	**82.5**	75.6	33.3	20.0	20.0	20.0
	Dominance Display	80.0	75.0	75.0	**82.5**	35.0	30.0	37.5	64.4
Multi-Agent Social Strategy	Helping Based on Visual Perspective	42.2	42.2	**84.4**	53.3	24.4	40.0	17.8	33.3
Overall	-	56.4	52.2	**79.6**	72.4	30.8	39.2	27.6	48.3

4.1. Experiment Setup

We evaluate our benchmark on a diverse set of state-of-the-art text-to-video generation systems, covering both proprietary and commercial-grade models. Concretely, for closed-source model we include four representative models: Sora2pro[36], Kling2.5turbo[21], Veo3.1[13], Hailuo02 Standard(Hailuo02-S)[15]. For open-source model, we select three representative models: Hunyuan-Video[23], LTX-1.0[14], LongCat-Video[45].We present our experiments in the following sections. For evaluation, we use Gemini 2.5 Pro [6] as our vision–language model (VLM) judge.

4.2. Evaluation Results

Table 2 summarizes the performance of representative text-to-video models across the 15 socially grounded tasks, grouped into six major dimensions of social reasoning.

First, Sora2-Pro and Veo-3.1 exhibit a clear advantage across nearly all task categories, achieving overall pass rates of 79.6% and 72.4%, respectively. Their strengths are particularly pronounced in tasks involving goal understanding, joint attention, and prosocial behavior, where both models exceed 80% on most sub-tasks. These results suggest that top-tier proprietary systems already possess strong implicit priors for human motion causality, gaze direction, and intention-driven interactions, even without explicit cue engineering. In contrast, Hailuo02-S and Kling2.5-Turbo show substantially weaker reasoning ability, with overall scores of 56.4% and 52.2%. These models struggle especially with tasks that require coordinated multi-agent behavior (e.g., Leader–Follower Coordination) or abstract social inference (e.g., Helping Based on Visual Perspective), exhibiting failure rates over 50%. Their performance improves noticeably when explicit cues are available (e.g., Pointing Comprehension), indicating a higher reliance on surface-level visual signals.

A significant performance gap emerges when comparing these proprietary systems to their open-source counterparts. The evaluated open-source models—Longcat-Video, HunyuanVideo, and LTX-1.0—operate at a substantially lower performance level across nearly all dimensions. Their struggles are particularly acute in tasks requiring complex causal or belief-state reasoning. Among this group, Longcat-Video shows comparatively stronger, though still limited, capabilities, while HunyuanVideo and LTX-1.0 lag further behind. This disparity underscores the current chasm in sophisticated social reasoning between state-of-the-art proprietary models and the broader open-source ecosystem, likely reflecting differences in model scale, training data, and architectural design for capturing complex agent interactions.

4.3. Verification of the Agent-Based Generation

Validation of Prompt Quality Across Pipeline Stages. A key contribution of our benchmark is the agent-based pipeline that transforms abstract psychological paradigms into concrete, video-ready prompts. To validate the quality of the constructed prompts, we conduct a human evaluation comparing three stages of the generation process.

We test prompts generated at three stages of the pipeline: (1) without any conceptual understanding (“No Understanding”), (2) with Experiment Understanding followed by Prompt Synthesis (“Understanding + Synthesis”), and (3) the full pipeline including Critic Agent revision (“Full”).

Table 3. Human pass rates (%) for prompts from pipeline stages: (1) No Understanding, (2) +Synthesis, (3) Full pipeline.

Dimension	**No Und.**	**+Synth.**	**Full**
Goal Directed Action	68.1	76.5	87.5
Joint Attention & Perspective	66.5	75.2	86.3
Social Coordination	67.2	74.5	86.5
Emotion & Prosocial	68.3	78.3	88.2
Social Norms & Spacing	66.4	77.2	87.2
Multi-Agent Strategy	65.6	73.5	85.6
Mental State Reasoning	65.8	76.1	87.2
Average	**66.8**	**75.9**	**86.9**

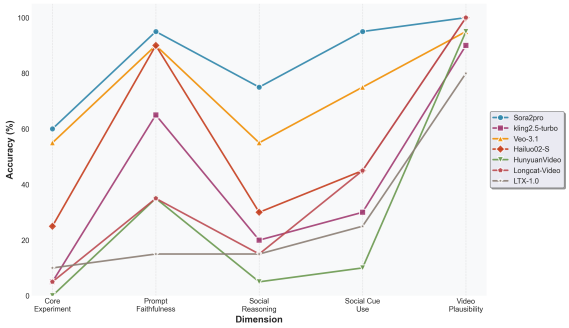
Table 4. Average pass rates (%) under different difficulty levels for four closed-source models. The results reflect how cue-based difficulty modulation affects each model's performance.

Models	Easy	Mid	Hard
Sora2pro	73.8	84.8	79.4
Veo3.1	66.6	74.4	75.8
Hailuo02-S	62.6	56.8	49.8
Kling2.5turbo	58.0	54.0	44.6

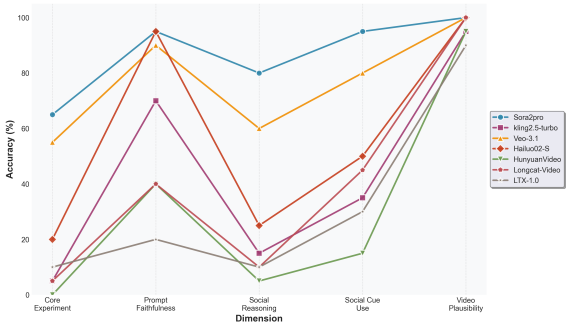
Human judges simply decide whether a prompt correctly expresses the intended reasoning construct while remaining descriptively neutral and free of answer leakage.

As shown in Table 3, incorporating conceptual understanding substantially improves prompt validity across all seven social reasoning dimensions. Pass rates rise from roughly 67.7% without understanding to over 75.8% once the test point is explicitly distilled, and further to above 86.7% when Critic Agent refinement is applied. These results demonstrate that the reasoning-aware generation stage—and its subsequent critic-driven correction—are both essential for producing prompts that are theoretically faithful and suitable for evaluating social reasoning in video generation models.

Cue-Controllability and Difficulty Ordering. We further validated that the easy, medium, and hard variants produced by the Critic Agent form a meaningful difficulty hierarchy that modulates the reasoning challenge faced by different video generation systems. As shown in Table 4, the cue-based variants yield a clear difficulty ordering for *Hailuo02-S* and *Kling2.5-Turbo* (Easy $>$ Medium $>$ Hard), confirming that richer social cues effectively facilitate reasoning in models with weaker inference capacity. However, *Sora2-Pro* and *Veo3.1* exhibit a reversed trend: their performance peaks at the medium or hard conditions despite reduced visual cues. We attribute this to their stronger intrinsic social reasoning capability—these models can infer social intent and causal structure even under minimal information, and additional cues may introduce conflicting or redundant signals that cannot be perfectly realized within a 5–10s clip, leading to penalties on other dimensions such as *Prompt Faithfulness* or *Social Cue Use*. In contrast, the weaker models benefit directly from explicit cues, showing a monotonic improvement when the task becomes more guided. Together, these results confirm that our cue-based difficulty design not only controls the inferential complexity of each prompt but also reveals distinct reasoning regimes across models: high-performing systems demonstrate robustness under cue sparsity, whereas lower-capacity models rely heavily on external cue enrichment.



(a) Agent evaluation results



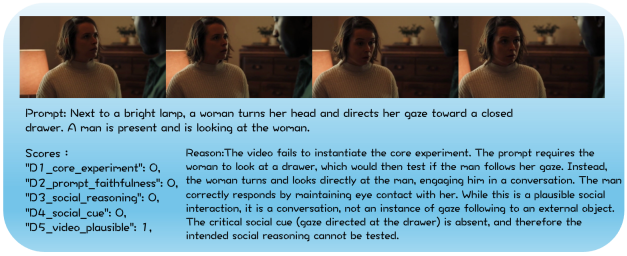
(b) Human evaluation results

Figure 3. Comparison between automated and human evaluation. (a) Scores from the agent-based VLM Judge. (b) Scores from human annotators using the same five-dimensional rubric. The two profiles exhibit closely matched trends across dimensions and models.

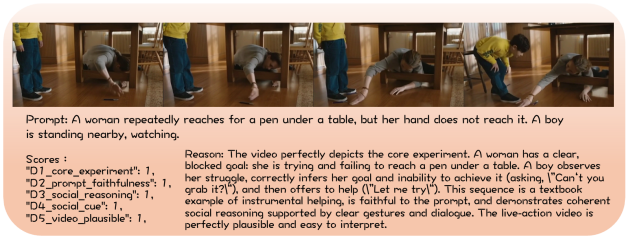
4.4. Validation of the Agent-Based Evaluation

Our benchmark relies on an automated VLM Judge to assess social coherence, task correctness, and causal consistency in generated videos. To ensure that this automatic evaluation is trustworthy, we perform several analyses examining its reliability and alignment with human reasoning.

Human Alignment To validate our automated evaluation protocol, we utilize Gemini 2.5 Pro as our Vision-Language Model (VLM)-based judge. As exhaustive human annota-



(a)



(b)

Figure 4. Representative qualitative cases of the agent-based evaluator. Each panel shows sampled video frames, the generation prompt, the five-dimensional scores, and the model's natural-language justification.

tion across all models is prohibitively expensive, we adopt a stratified sampling strategy. We randomly sampled 20 generated videos per model, creating a 160-clip test set to compare the VLM's judgments against human ratings across all five dimensions. As shown in Figure 3, our comparative analysis reveals a high degree of alignment in the relative performance trends between the VLM judge (Fig.3a) and human annotators (Fig.3b). The relative difficulty of the evaluation dimensions remains consistent for both. However, we identify a systematic divergence in scoring thresholds. Human annotators are more lenient on perceptual dimensions like *Prompt Faithfulness (D2)*, *Social Cue Use (D4)*, and *Video Plausibility (D5)*, where their pass rates approach ceiling. This suggests humans prioritize sufficient visual clarity for interpretability over pixel-perfect generation. Conversely, on reasoning-intensive dimensions—*Core Experiment (D1)* and *Social Reasoning (D3)*—human raters are markedly stricter. They demand that the complete causal and logical structure of the psychological paradigm be correctly instantiated, penalizing even minor deviations from the intended social logic. This intuitive bias reflects a tolerance for surface-level imperfections but an intolerance for logical flaws.

Qualitative Case Analysis To illustrate the practical application of our framework, Figure 4 presents two representative cases that showcase the VLM judge's decision-making process.

The first case (Fig. 4a) demonstrates a critical failure mode. The prompt specifies a gaze-following experiment where a woman's gaze should guide a man's attention to a drawer. The resulting video, however, shows a plausible but incorrect interaction—a direct conversation with mutual eye contact. Our agent correctly diagnoses this mismatch: it recognizes the video's visual quality (D5: Video Plausibility = 1) but correctly assigns zeros to the other four dimensions because the core experiment was not performed, the prompt was not followed, and the specific social cues for gaze-following were absent.

The second case (Fig. 4b) provides a textbook example of a successful generation. Tasked with showing instrumental helping, the video portrays a boy noticing a woman's repeated failed attempts to reach a pen and subsequently intervening to help her. This sequence perfectly aligns with the prompt and the intended social logic. As a result, the VLM judge awards a full score across all five dimensions, confirming that the model successfully rendered the prompt's narrative, the core psychological concept, coherent agent reasoning, and effective social cues within a plausible scene. These cases validate our agent's capacity to not only reward faithful and logically sound generations but also to penalize subtle yet critical failures where the underlying experimental paradigm is violated, even if the resulting video appears socially coherent on the surface.

5. Conclusion

We present the first benchmark dedicated to evaluating *social reasoning* in video generation. Unlike existing evaluations that focus primarily on perceptual or physical dimensions, our benchmark targets the causal, intentional, and socially grounded behaviors that underlie human interaction. By grounding the benchmark in eight well-established components of social cognition and operationalizing them through thirty classic psychological experiments, we provide a theoretically principled and interpretable foundation for assessing social reasoning in generative models.To enable scalable and training-free benchmark construction, we introduce a four-agent pipeline that transforms abstract psychological paradigms into validated, difficulty-controlled, video-ready prompts and performs automatic evaluation through a vision–language model judge. Our analyses demonstrate that both generation-side agents and evaluation-side agents exhibit strong alignment with human reasoning, enabling reliable large-scale assessment without manual annotation. Experiments across eight state-of-the-art video generation models reveal a substantial gap between perceptual realism and social coherence: while leading proprietary systems show emerging competence in goal understanding, joint attention, and prosocial behavior, even the strongest models fail systematically on belief-based inference, subtle cue integration, and multi-agent coordination.

References

[1] Roger Bakeman and Lauren B. Adamson. Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, 55(4):1278–1289, 1984. 5

[2] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46, 1985. 2, 5

[3] C. Daniel Batson. *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991. 5

[4] Dana R. Carney, Judith A. Hall, and Lynn S. LeBeau. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–123, 2005. 5

[5] Yongfan Chen, Xiuwen Zhu, and Tianyu Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction. *arXiv preprint arXiv:2502.05503*, 2025. 3

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[7] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972. 5

[8] Nancy Eisenberg and Richard A. Fabes. Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion*, 14(2):131–149, 1990. 5

[9] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. 5

[10] John H. Flavell. The development of knowledge about visual perception. In *Nebraska Symposium on Motivation*, pages 43–76. University of Nebraska Press, Lincoln, NE, 1978. 5

[11] C. Keith Friesen and Alan Kingstone. The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5(3):490–495, 1998. 5

[12] György Gergely and Gergely Csibra. Goal attribution to agents by 6.5-month-old infants. In *The cognitive basis of imitation*, pages 228–249. Blackwell Publishers, 2002. 2, 5

[13] Google. Veo 3 and 3.1. https://aistudio.google.com/models/veo-3, 2025. 6

[14] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 6

[15] Hailuo AI. Hailuo. https://hailuoai.video/, 2025. 1, 6

[16] Edward T. Hall. *The Hidden Dimension*. Doubleday, 1966. 2, 5

[17] Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–100, 1993. 5

[18] Dirk Helbing and Peter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995. 5

[19] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 3

[20] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 3

[21] Kling AI. Kling 2.5 turbo. https://app.klingai.com/global/release-notes/2025-09-19, 2025. 1, 6

[22] Wolfgang Köhler. *The Mentality of Apes*. Harcourt, Brace, New York, 1925. 5

[23] W Kong, Q Tian, Z Zhang, R Min, Z Dai, J Zhou, J Xiong, X Li, B Wu, J Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models, 2025. *URL https://arxiv. org/abs/2412.03603*. 6

[24] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169, 2023. 3

[25] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. T2v-compbench: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 2

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 2

[27] Jeffrey J. Lockman. The development of detour ability during infancy. *Child Development*, 55(2):482–491, 1984. 5

[28] Bertram F. Malle. The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33(3):101–121, 1997. 5

[29] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025. 3

[30] Kerry L. Marsh, Michael J. Richardson, and Richard C. Schmidt. Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, 1(2): 320–339, 2009. 5

[31] Zenaida S. Masangkay, Kevin A. McCluskey, Clinton W. McIntyre, Judith Sims-Knight, Billy E. Vaughn, and John H. Flavell. The early development of inferences about the visual percepts of others. *Child Development*, 45(2):357–366, 1974. 5

[32] Andrew N. Meltzoff. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850, 1995. 2, 5

[33] Author Miller. Queue behavior, 2002. 5

[34] Henrike Moll and Michael Tomasello. How 14- and 18-month-olds know what others have experienced. *Developmental Psychology*, 43(2):309–317, 2007. 5

[35] Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Rˆ 3-vqa:" read the room" by video social reasoning. *arXiv preprint arXiv:2505.04147*, 2025. 3

[36] OpenAI. Sora 2. https://openai.com/index/sora-2/, 2025. 1, 6

[37] Wenshuo Peng, Kaipeng Zhang, Yue Yang, Hao Zhang, and Yu Qiao. Data adaptive traceback for vision-language foundation models in image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4506–4514, 2024. 3

[38] Wenshuo Peng, Kaipeng Zhang, and Sai Qian Zhang. T3m: Text guided 3d human motion synthesis from speech. *arXiv preprint arXiv:2408.12885*, 2024. 3

[39] Josef Perner, Susan R. Leekam, and Heinz Wimmer. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5 (2):125–137, 1987. 5

[40] Chris Pratt and Peter Bryant. Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, 61(4):973–982, 1990. 5

[41] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978. 2

[42] Hannes Rakoczy, Felix Warneken, and Michael Tomasello. Taking fiction seriously: Young children understand the normative structure of joint pretend games. *Developmental Psychology*, 44(4):1195–1201, 2008. 5

[43] Amanda Seed and Richard W. Byrne. Animal tool-use. *Current Biology*, 20(23):R1032–R1039, 2010. 5

[44] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[45] Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025. 6

[46] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691, 2005. 2, 5

[47] Michael Tomasello, Malinda Carpenter, and Ulf Liszkowski. A new look at infant pointing. *Child Development*, 78(3): 705–722, 2007. 5

[48] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023. 3

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 3

[50] Felix Warneken and Michael Tomasello. Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765): 1301–1303, 2006. 2, 5

[51] Felix Warneken, Frances Chen, and Michael Tomasello. Cooperative activities in young children and chimpanzees. *Child Development*, 77(3):640–663, 2006. 5

[52] Andrew Whiten and Richard W. Byrne. Machiavellian intelligence. In *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*, pages 1–9. Clarendon Press/Oxford University Press, 1988. 2, 5

[53] Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*, 2024. 3

[54] 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*, pages 8807–8817, 2019. 3

[55] Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025. 3

[56] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 3