SMORE: MINIMAL-PAIR VIDEOS FOR SOCIAL, MORAL AND RATIONAL REASONING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding human behavior involves interpreting actions in relation to agents' beliefs, goals and social norms. Small differences in behavior can change these interpretations even when the surrounding context remains similar. To test whether multimodal large language models (MLLMs) reliably distinguish such cases, we introduce SMORE, a minimal-pair benchmark for evaluating **S**ocial, **MO**ral, and **R**ational reasoning. It comprises 427 video pairs (854 videos) and 1,281 evaluation questions spanning belief-conditioned rationality, helping and hindering and judgements of intentionality, responsibility and norm violations. Videos within each pair share the same actors, environment, and overall interaction while differing in a single socially-relevant factor that changes the target judgement. This design probes whether model judgements track relevant behavioral differences under closely matched conditions. We evaluate models under two settings. First, we have paired-video evaluation, where both videos from a minimal pair are presented jointly to measure relative discrimination performance. Second, we have single-video evaluation, where each video is evaluated independently to measure absolute performance. From the second setting, we additionally compute Joint-Pair Accuracy, which gives credit only when both videos in a pair are individually answered correctly. On SMORE, humans reach 96% Joint-Pair Accuracy while the best model, Gemini-3.5-Flash, reaches only 63.2%. These results suggest that, despite strong performance on video action understanding, current MLLMs still fall substantially short of humans in reasoning about the subtle differences in human behavior that determine social, moral, and rational judgments. The dataset is available at `https://anonymous.4open.science/r/SMORE-2B49`.

1 INTRODUCTION

Consider a simple everyday interaction: Alice is tidying up her living room and walks toward a cabinet, carrying a book she intends to put away. The cabinet door is slightly open, and David is standing next to it. Just before she reaches it, David pulls the cabinet door open, and Alice places the book inside without breaking stride. Now imagine an almost identical interaction: the same room, the same two people, the same book, and the same goal. This time, as Alice approaches, David pushes the cabinet door shut, leaving her standing in front of it with the book still in her hands. In which scene is David helping Alice, and in which is he hindering her? Did he act deliberately or by accident? Was his action sensible given what he believed? These are questions human observers ask and answer effortlessly in everyday life. Yet answering them requires moving beyond recognizing what people are doing to reasoning about why they act and how their actions affect others, interpreting behavior in light of goals, beliefs and intentions.

This ability is central to understanding social behavior, yet evaluating it poses a particular challenge: a model may answer an individual example correctly by relying on contextual or superficial cues without necessarily identifying the socially relevant difference that determines the judgment. A strong evaluation should therefore test not only whether a model produces the correct judgment, but whether that judgment changes appropriately when the relevant behavior changes while the surrounding context remains fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Benchmark	#Items	#QA	Video	Constructs Evaluated				Evaluation Format
				Belief-cond. Rationality	Social Reasoning	Moral Reasoning	Minimally Diff. Videos	
AGENT (Shu et al., 2021) (ICML)	8,400	3,360†	✓	∼	✗	✗	∼	VoE
BIB (Gandhi et al., 2021) (NeurIPS)	10,000	5,000†	✓	∼	✗	✗	∼	VoE
MuMA-ToM (Shi et al., 2025) (AAAI)	225	900	✓	✗	✓	✗	✗	SV
SoMi-ToM (Fan et al., 2025) (NeurIPS)	35	1,225	✓	✗	✓	✗	✗	SV
MoMentS (Villa-Cueva et al., 2025) (EMNLP)	168	2,335	✓	✗	✓	✗	✗	SV
SIV-Bench (Kong et al., 2026) (ACL)	2,792	5,455	✓	✗	✓	✗	✗	SV
R^3-Bench (Niu et al., 2026) (ICLR)	316	5,100	✓	✗	✓	✗	✗	SV
SocialReasonBench (Huang et al., 2026) (EMNLP)	532	532	✓	✗	✓	✗	✗	SV
VIVA (Hu et al., 2024) (EMNLP)	1,240	1,240	✗	✗	✗	✓	✗	SI
M^3oralBench (Yan et al., 2024) (PR)	1,160	4,640	✗	✗	✗	✓	✗	SI
MORALISE (Lin et al., 2026) (ICML)	2,481	2,481	✗	✗	✗	✓	✗	SI
SMORE (Ours)	**854**	**1,281**	✓	✓	✓	✓	✓	**SV+J+P**

Table 1: **Comparison with prior benchmarks targeting related reasoning capabilities.** SMORE jointly evaluates belief-conditioned rationality, goal-directed social reasoning, and moral reasoning using controlled minimal-pair videos. #Items denotes the number of videos/images and #QA the number of evaluation questions. ∼ denotes partial coverage: AGENT and BIB evaluate rational or goal-directed behavior in simplified violation-of-expectation settings without explicitly modeling the agent's belief state, where expected/unexpected video pairs are compared jointly rather than evaluated independently as minimally differing videos; for these benchmarks, † denotes the number of evaluation trials formed from such pairs. ***VoE***: violation of expectation; ***SV***: single-video MC-QA; ***SI***: single-image MC-QA; ***SV+J+P***: single-video MC-QA as in prior work, additionally scored with Joint-Pair Accuracy, and paired-video MC-QA.

We address this problem using *counterfactual minimal video pairs*: closely matched videos in which the broader interaction is held fixed while a targeted change alters the correct interpretation. Prior work has shown that controlled video comparisons can reveal substantial weaknesses that conventional benchmark evaluations obscure. By constructing conflicting, counterfactual, or minimally differing videos, these benchmarks limit models' ability to succeed through static visual cues, language priors, or other superficial correlations. In doing so, they expose failures in temporal and physical understanding that can remain hidden despite apparently strong performance on standard video-understanding tasks (Liu et al., 2024; Zhang et al., 2024; Krojer et al., 2025).

However, counterfactual minimal-pair evaluation and aspects of social behavior understanding have largely been studied separately. Prior work has studied instrumental rationality and goal-directed behavior through violation-of-expectation paradigms (Shu et al., 2021; Gandhi et al., 2021); social and multi-agent reasoning through benchmarks targeting beliefs, goals, intentions, and interactions (Shi et al., 2025; Fan et al., 2025; Villa-Cueva et al., 2025; Kong et al., 2026; Niu et al., 2026; Huang et al., 2026); and moral reasoning through benchmarks evaluating moral judgments and related distinctions (Hu et al., 2024; Yan et al., 2024; Lin et al., 2026). As summarized in Table 1, however, these capabilities have not been jointly studied through controlled minimally differing videos. Existing work has therefore examined these aspects of social understanding without the controlled comparisons needed to test whether model judgments respond to the specific behavioral distinction that determines the interpretation.

To address this gap, we introduce SMORE, a controlled minimal-pair video benchmark for social behavior understanding. We evaluate three aspects through minimal video pairs: *belief-conditioned rationality*, whether an action is rational given what an agent believes; *goal-directed social reasoning*, whether one agent's behavior helps or hinders another's goal; and *moral reasoning*, spanning intentionality, responsibility taking, and moral–conventional transgression. To our knowledge, SMORE is the first benchmark to evaluate these aspects of social behavior understanding in MLLMs through controlled minimal video pairs.

The minimal-pair structure of SMORE also changes how models can be evaluated. Most prior benchmarks score each video or image independently, which measures whether individual judgments are correct. However, a model that relies on context shared across the two videos in a minimal pair, rather than on the behavior that differs between them, may give the same judgment to both.

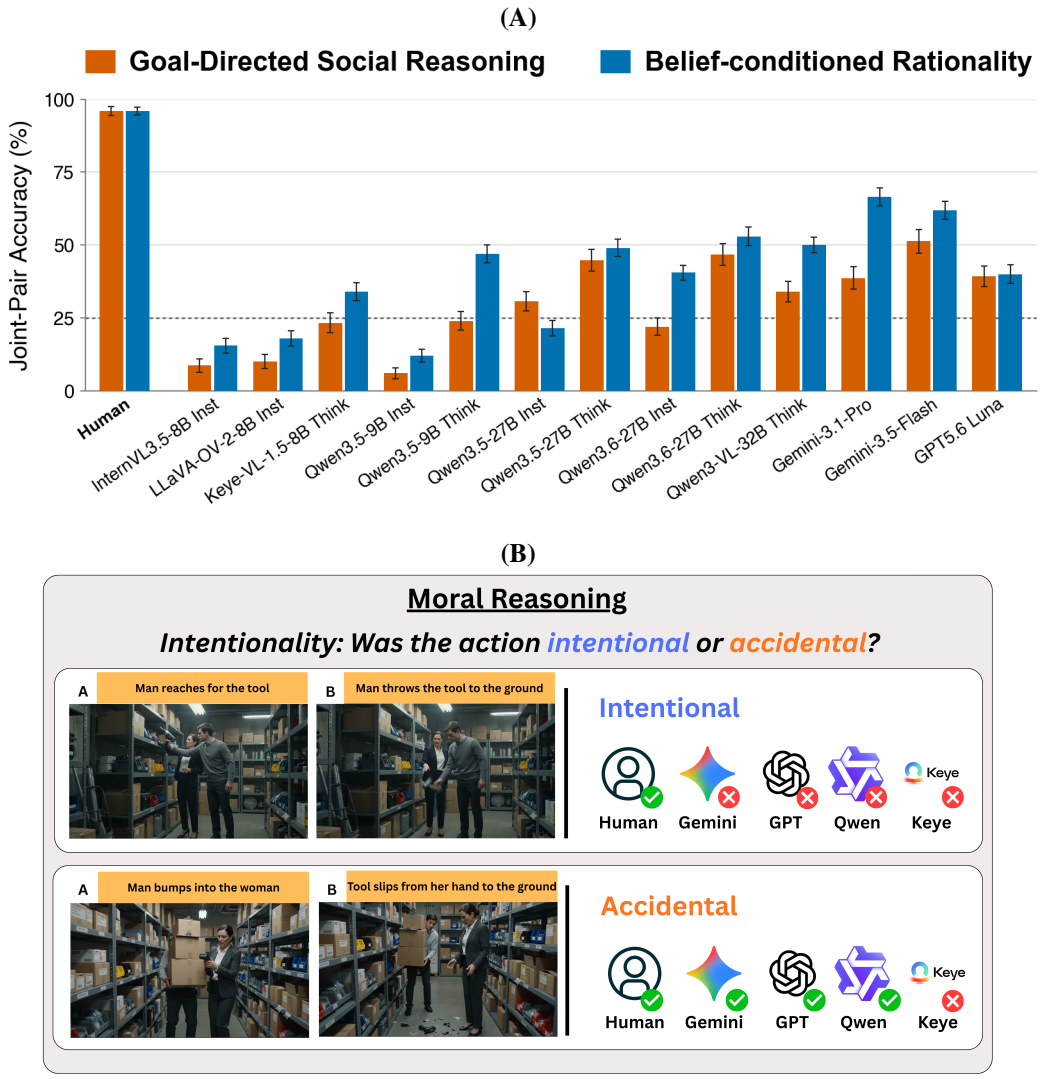

Figure 1: **A substantial human–MLLM gap on SMORE. (A)** Joint-Pair Accuracy for belief-conditioned rationality and goal-directed social reasoning across all evaluated models. Joint-Pair Accuracy requires both independently evaluated videos of a minimal pair to be classified correctly (chance: 25%, dashed line). Error bars show ± 1 bootstrap standard error from the 10,000 bootstrap resamples of minimal pairs samples. Humans remain near ceiling, while every evaluated MLLM falls substantially below human performance, with several models near or below chance. **(B)** SMORE additionally evaluates moral reasoning. Shown here is a minimal pair testing whether an action is intentional or accidental. Here, three of the four models judge both videos as accidental, failing to recognize the minimal behavioral difference that determines the label.

Because the two members have opposing labels, such a model still receives partial credit when accuracy is measured over individual videos (which we refer to as Single-video Accuracy). We therefore also report Joint-Pair Accuracy, which gives credit only when both videos in a pair are classified correctly, that is, only when the model's judgment changes with the behavior, and in the right direction. Violation-of-expectation benchmarks take a different approach, comparing expected and unexpected videos jointly without requiring an absolute judgment of either. Such a comparison is informative: judging a single video requires an implicit threshold for how much evidence is enough, whereas seeing both videos together supplies the reference directly. Our third metric, Paired Accuracy, follows this approach: both videos are shown together and the model must identify which one exhibits the target property. A large gap between Paired and Single-video Accuracy can indicate that a model detects the behavioral difference within a pair but cannot reliably place an absolute judgment on each video individually.



Figure 2: **Dataset examples from SMORE.** We provide keyframes from both videos in each minimal pair. The captions above the frames are illustrative only and are not shown. Goal-directed social reasoning and Belief conditioned rationality show the *Open/Close* and *False Belief* subclasses, respectively. The questions correspond to the single-video evaluation setting, in which each video is evaluated independently. Across all examples, the ground-truth answer is Option 1 (blue) for Video A and Option 2 (orange) for Video B. In the paired-video evaluation setting, the questions are modified to jointly present both videos and ask the model to distinguish between them.

Our contributions are fourfold: (1) we introduce SMORE, a controlled minimal-pair video benchmark spanning belief-conditioned rationality, goal-directed social reasoning, and moral reasoning; (2) we evaluate models using complementary paired-video and single-video protocols to measure relative and absolute performance, and use Joint-Pair Accuracy to measure whether both independently evaluated members of a minimal pair are classified correctly; (3) we probe model representations to examine where these judgments become linearly decodable and whether they remain linearly decodable even when models make incorrect predictions; and (4) we benchmark state-of-the-art MLLMs alongside human judgments, revealing a substantial gap between models and humans (Figure 1).

2 SMORE Benchmark

2.1 Benchmark construction

Each example consists of two videos that preserve actors, objects, environment, camera viewpoint, initial conditions, and task context while modifying a single judgment-relevant variable. The benchmark contains **427 minimal pairs (854 videos) and 1,281 questions** across three reasoning domains.

Belief-conditioned rationality. Rational behavior involves choosing actions that efficiently achieve an agent's goal given its beliefs about the world (Gergely & Csibra, 2003; Baker et al., 2009). We construct four controls of 50 pairs each (200 pairs total). In *True Belief*, the rational agent follows a correct belief and succeeds, whereas the irrational agent acts against it and fails. In *False Belief*, the rational agent follows an incorrect belief and fails, whereas the irrational agent acts against it and

happens to succeed, thereby *decorrelating goal achievement from rationality*. In *Path Believed Safe*, taking the shortest path is rational and an unnecessary detour irrational; in *Path Believed Risky*, the rational agent instead takes a longer detour while the shorter direct path is irrational. These settings *decorrelate efficiency from rationality*.

Goal-directed social reasoning. Social evaluation depends on whether one agent facilitates or obstructs another agent's goal (Hamlin et al., 2007). We test three action classes. In *Find* (59 pairs), one agent asks another to find an object; the second agent either brings it closer, thereby helping, or moves it farther away, thereby hindering. In *Open/Close* (47 pairs), the first agent needs an open container to complete a household task, and the second agent either opens it or closes it mid-task. In *On/Off* (44 pairs), the first agent needs to use a device, and the second agent either switches it on or switches off one already in use.

Moral reasoning. We evaluate three components of moral cognition. *Intentionality* (25 pairs) distinguishes deliberate from accidental actions (Rizzo & Killen, 2014; Guglielmo, 2015); *Responsibility taking* (27 pairs) tests whether an agent acknowledges responsibility for a negative outcome it caused rather than deflecting it (Guglielmo, 2015); and *Moral vs. Conventional Transgressions* (25 pairs) distinguishes violations involving harm, rights, or property from context-dependent social rules or customs (White et al., 2017).

Video generation and minimal-pair control. We use two generation routes, matched to what each domain requires. Rationality and goal-directed social reasoning videos are procedurally generated in VirtualHome (Puig et al., 2018). Both videos of a pair are rendered from a single sampled configuration that fixes the apartment, agents, objects, camera pose, and initial scene state, and they differ only in one scripted action. Moral reasoning scenarios involve behaviors such as accidents, owning or deflecting a caused outcome, and norm violations, which fall outside a simulator's action vocabulary. We therefore generate them with Gemini Omni (Google DeepMind, 2026c). Both videos of a pair are conditioned on the same actor and location reference images, and on prompts that differ only in the judgment-relevant action. Details of both pipelines are in Appendix A.6.1 and A.6.2.

We validate minimal-pair control in two ways. In Qwen3.6-27B's visual representation space, videos within a pair are substantially closer than videos from different pairs (Appendix A.6.3). In a human study on a representative subset, annotators nearly always reported that the key difference between the two videos in a pair was the intended behavioral contrast (Appendix A.6.4).

2.2 Evaluation protocols

We evaluate models under two settings. In **paired-video evaluation**, both videos of a minimal pair are shown together and the model picks the one matching the target judgment; **Paired Accuracy** (P) is the fraction of pairs answered correctly. In **single-video evaluation**, each video is judged independently; **Single-video Accuracy** (S) is the fraction of videos classified correctly, and **Joint-Pair Accuracy** (J) is the fraction of pairs for which both videos are correct. Chance is 50% for P and S, and 25% for J.

3 Empirical Results on SMORE

We evaluate a range of state-of-the-art proprietary and open-source MLLMs: InternVL3.5 (Wang et al., 2025), LLaVA-OneVision-2 (An et al., 2026), Keye-VL-1.5 (Yang et al., 2025a), Qwen3.5 (Qwen Team, 2026a), Qwen3.6 (Qwen Team, 2026b), Qwen3-VL (Bai et al., 2025), Gemini-3.1-Pro (Google DeepMind, 2026a), Gemini-3.5-Flash (Google DeepMind, 2026b), and GPT-5.6 Luna (OpenAI, 2026). For the Qwen models, we evaluate both thinking and instruct configurations where supported. Figure 1 summarizes overall Joint-Pair Accuracy across all evaluated models, while Figure 3 shows Paired, Single-video, and Joint-Pair Accuracy across reasoning domains and subclasses for representative models. Complete results for all evaluated models are provided in Appendix Table 6.

3.1 Single-video accuracy masks poor pair-level accuracy

A model that relies on context shared across a minimal pair, rather than on the manipulated behavior, may give the same judgment to both videos. Since the two videos have opposing labels, such a prediction is correct for one and wrong for the other: it receives partial credit under S but none under

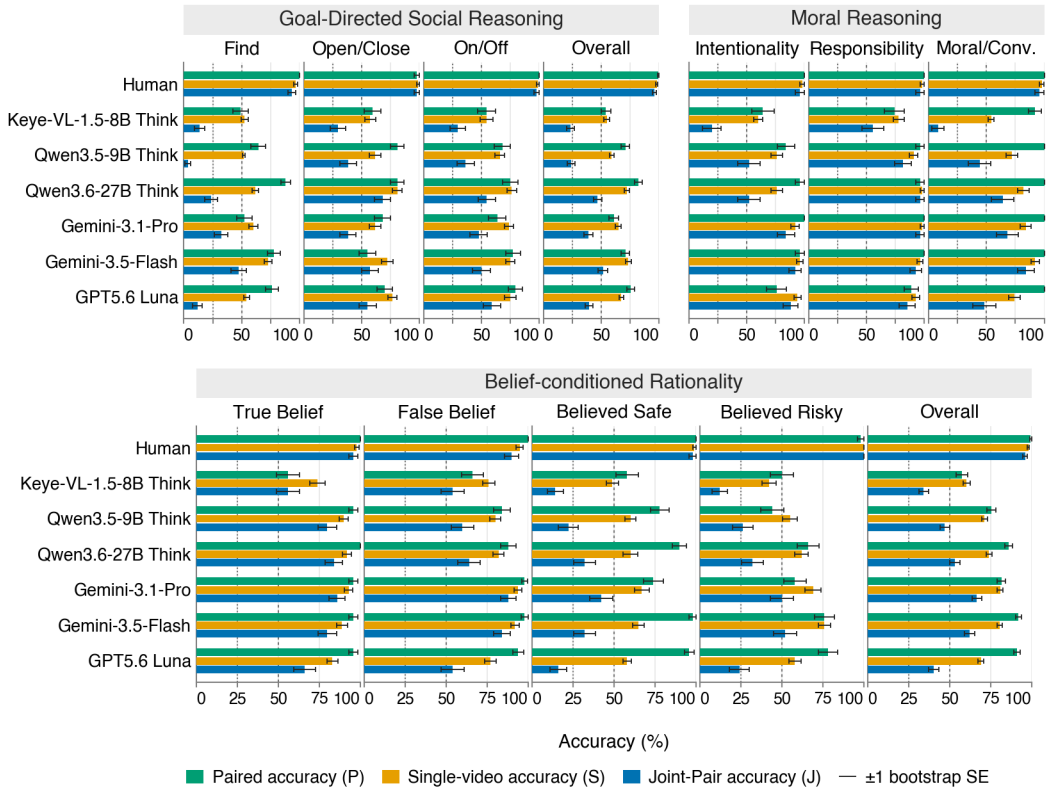

Figure 3: **Evaluation results on SMORE (%).** We report Paired accuracy (50% chance), Single-video accuracy (50% chance), and Joint-Pair Accuracy (25% chance), as defined in Section 2.2. Error bars show ± 1 bootstrap standard error from the 10,000 bootstrap resamples of minimal pairs samples. Human scores use three independent annotations per question from a pool of 65 annotators. Complete results for all evaluated models are reported in Appendix Table 6.

J. Comparing S and J therefore reveals whether apparent performance on individual videos translates into consistent judgments across the controlled behavioral contrast.

Humans show only a small gap between the two metrics: S and J differ by approximately two points on both Rationality (S = 98.0% vs. J = 96.0%) and Goal-Directed Social Reasoning (S = 98.0% vs. J = 96.0%). For models, the discrepancy is substantially larger. Even the proprietary models show sizeable differences between S and J. For example, Gemini-3.1-Pro drops from S = 80.8% to J = 66.5% on Rationality, and from S = 64.7% to J = 38.7% on Goal-Directed Social Reasoning.

Notably, some models achieve S at or above the 50% chance level while falling below the 25% chance level under J across multiple reasoning domains. On FIND, Qwen3.5-9B-Think obtains S = 51.7%, which would be consistent with random guessing. A model that guessed at random, however, would reach a J of about 25%; Qwen3.5-9B-Think reaches only J = 3.4%. Its errors are therefore not random but systematic: it assigns the same judgment to both videos in nearly every pair, regardless of whether the agent helps or hinders. Thus S alone can hide qualitatively different behavior, and only J reveals whether a model's judgments are sensitive to the behavioral change within the minimal pair.

3.2 COMPARING BOTH VIDEOS CAN MAKE THE CONTRAST EASIER

The paired and single-video settings pose the same underlying question in two different forms. In the single-video setting, a model must decide on its own whether a video shows the target property, for instance whether an action is rational or helpful. This requires an implicit threshold for how much evidence is enough. In the paired setting, the model only needs to decide which of the two

videos shows the property; the comparison itself supplies the reference, and no threshold is needed. Some advantage for P is expected even for an ideal observer, since two videos provide more evidence than one. A large gap between P and S, however, indicates that a model can detect the behavioral difference within a pair but cannot reliably place an absolute judgment on each video.

Several subdomains show exactly this dissociation. On Believed Safe, GPT-5.6 Luna identifies the correct video in 96% of pairs, yet classifies only 58% of the same videos correctly when they are judged independently. Gemini-3.5-Flash shows the same pattern (P = 98%, S = 65%). On Moral/Conventional Transgressions, five of the six models reach P = 100%, yet their S ranges only from 72% to 92%. On FIND, Qwen3.6-27B-Think reaches P = 88% but S = 62%. Humans, by contrast, are near ceiling in both settings.

3.3 MODELS SHOW AN AGENT-FAVOURABLE PREDICTION ASYMMETRY

We next ask whether single-video errors are symmetric with respect to the direction of the judgment. Each minimal pair contains one agent-favourable and one agent-unfavourable video, so the classes are balanced by construction. Figure 4 reports class-conditional accuracy after aggregating subclasses within each domain. In each domain, the favourable judgment is the one that reflects better on the agent: *Rational* for Belief-conditioned Rationality, *Help* for Goal-directed Social Reasoning, and *Unintentional*, *Takes Responsibility*, or *Conventional Transgression* for the three Moral subclasses.

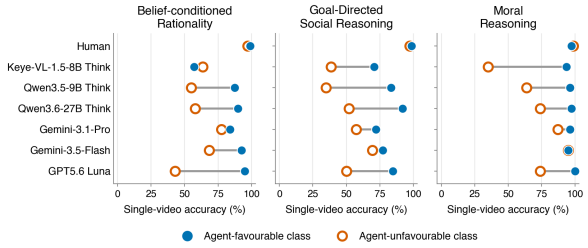

Figure 4: Class-conditional single-video accuracy for representative models. Filled markers denote agent-favourable cases and open markers denote agent-unfavourable cases. Models show a consistent advantage on agent-favourable examples, while human performance is nearly symmetric. Full results across all evaluated models are reported in Appendix Table 7.

The asymmetry is pronounced across all three domains. In Belief-conditioned Rationality, GPT5.6 Luna achieves 95.0% accuracy on the *Rational* class but only 43.0% on the *Irrational* class. In Goal-directed Social Reasoning, Qwen3.5-9B-Think achieves 83.3% on the *Help* class but only 34.7% on *Hinder* class. In Moral Reasoning, Keye-VL-1.5-8B-Think achieves 93.5% accuracy on agent-favourable cases but only 35.1% on agent-unfavourable cases. In contrast, human performance is nearly symmetric across the two classes in all domains. Overall, models are much more accurate when the correct judgment is favourable to the agent than when it is unfavourable.

3.4 GOAL ACHIEVEMENT AND PATH EFFICIENCY INFLUENCE RATIONALITY JUDGMENTS

Our Rationality conditions separate rational action from two shortcuts: goal achievement and path efficiency. In True Belief the rational action succeeds and the irrational one fails; False Belief reverses this, so the rational agent fails and the irrational one happens to succeed. Believed Safe makes the rational action the shorter direct path; Believed Risky requires the longer route.

We observe a consistent True- to False-Belief drop across all seven Qwen variants (Appendix Table 6): each obtains lower Paired accuracy on False Belief than on True Belief. For example, Qwen3.6-27B-Think falls from 100% to 88%, and Qwen3.5-9B-Inst falls from 92% to 34%. This drop suggests the Qwen models partly rely on goal completion when judging rationality (their reasoning traces support this: traces of incorrect predictions more often judge the action by its outcome; see Appendix A.9). The proprietary models show little or no such drop (e.g., Gemini-3.5-Flash rises from 96% to 98%). Because False Belief decorrelates goal achievement from rationality, such high accuracy there is only possible if models use the agent's stated belief rather than the observed outcome.

The effect of path efficiency is more widespread. Six of the seven Qwen variants obtain lower Paired accuracy on Believed Risky than on Believed Safe (Appendix Table 6). The proprietary models show the same direction of change: Gemini-3.1-Pro drops from 74% to 58%, Gemini-3.5-Flash from 98% to 76%, and GPT5.6 Luna from 96% to 78%. Thus, even when both videos are presented jointly, models generally perform better when the rational action is also the shorter route.

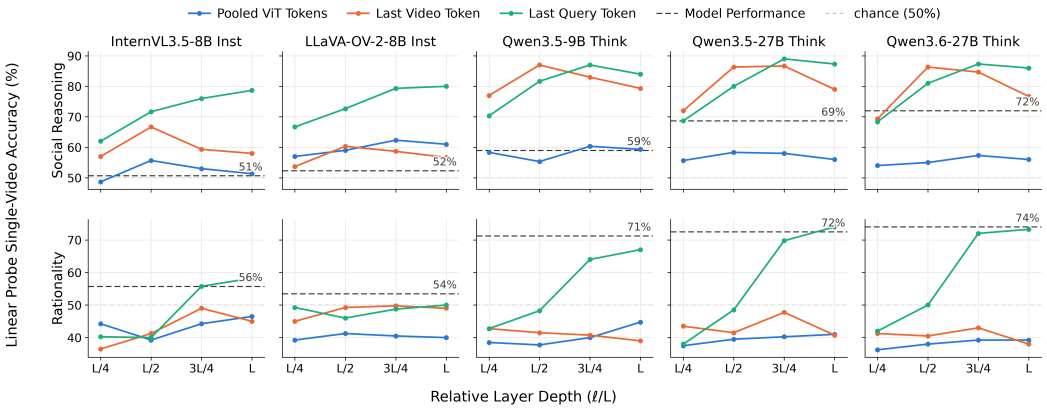

Figure 5: **Linear decodability across representations and model depth.** Probe accuracy for Goal-Directed Social Reasoning (top) and Belief-Conditioned Rationality (bottom) across five models, for the three features described in Section 4. Relative depth (ℓ/L) is measured within the vision encoder for Pooled ViT Tokens and within the language model otherwise. The light dashed line marks chance (50%); the dark dashed line marks the model's single-video accuracy.

4 Representation Analysis

The results above show current MLLMs often fail to track the behavioral changes isolated by SMORE. However, incorrect prediction does not reveal whether the judgment-relevant distinction can still be linearly decoded from the model's internal representations. We therefore complement the behavioral evaluation with a representation analysis using linear probes. The probe target is the benchmark label: Help versus Hinder for goal-directed social reasoning, and Rational versus Irrational for belief-conditioned rationality. This lets us ask two questions: *(1) where these judgments become linearly decodable, and (2) whether they remain decodable on incorrectly answered examples.*

Linear Probing setup. For five open-source models (InternVL3.5-8B Inst, LLaVA-OV-2-8B Inst, Qwen3.5-9B Think, Qwen3.5-27B Think, and Qwen3.6-27B Think), we train ℓ_2-regularized logistic-regression probes (Alain & Bengio, 2016; Hewitt & Liang, 2019) to predict the benchmark label from three pre-generation representations: mean-pooled vision-encoder patch states (*Pooled ViT Tokens*), the language-decoder hidden state at the final position of the video-token sequence (*Last Video Token*), and the language-decoder hidden state at the final position of the task query, where both the video and query have been jointly contextualized (*Last Query Token*). Each feature is extracted at depths $\{L/4, L/2, 3L/4, L\}$ of its component. Belief-conditioned rationality and goal-directed social reasoning are probed separately, with a distinct probe trained for each feature and layer. We use five-fold cross-validation with both videos of a minimal pair assigned to the same fold, so no probe is tested on a video whose counterpart it saw in training; classes are balanced, so chance is 50%. The full setup, along with results for the Moral Reasoning domain, is given in Appendix A.11.

4.1 Where do the judgments become linearly decodable?

Figure 5 shows different representational patterns for Goal-Directed Social Reasoning and Belief-Conditioned Rationality. For Goal-Directed Social Reasoning, Help/Hinder is only moderately decodable from Pooled ViT Tokens, but becomes substantially more decodable from the language-model features. Last Query Token probes reach approximately 80–89% accuracy across models, with the strongest results generally occurring at the last layers. Importantly, these probe accuracies are substantially higher than the corresponding single-video behavioral accuracy for all five models, indicating that the Help/Hinder distinction is more linearly recoverable from the internal representations than the models' own predictions would suggest.

Belief-Conditioned Rationality shows a different pattern. Pooled ViT Tokens and Last Video Token features remain close to chance across most models, and the Rational/Irrational distinction is not strongly or consistently linearly decodable from these representations. Although Last Query Token

Table 2: **Linear decodability conditioned on behavioral correctness.** Probe accuracy (%) from the final-layer Last Query Token, split by whether the model's own prediction is correct or wrong. MODEL CORRECT and Parentheses give the number of videos in each subset.

	GOAL-DIRECTED SOCIAL REASONING		BELIEF-CONDITIONED RATIONALITY	
MODEL	MODEL CORRECT	MODEL WRONG	MODEL CORRECT	MODEL WRONG
InternVL3.5-8B Inst	79.6 ($n = 152$)	77.7 ($n = 148$)	64.6 ($n = 223$)	50.3 ($n = 177$)
LLaVA-OV-2-8B Inst	83.4 ($n = 157$)	76.2 ($n = 143$)	50.5 ($n = 214$)	49.5 ($n = 186$)
Qwen3.5-9B Think	84.2 ($n = 177$)	83.7 ($n = 123$)	70.5 ($n = 285$)	58.3 ($n = 115$)
Qwen3.5-27B Think	89.3 ($n = 206$)	83.0 ($n = 94$)	80.3 ($n = 290$)	57.3 ($n = 110$)
Qwen3.6-27B Think	90.3 ($n = 216$)	75.0 ($n = 84$)	81.8 ($n = 296$)	49.0 ($n = 104$)

probes improve in the later layers of the three Qwen models, reaching approximately 67–74%, this improvement is much weaker for InternVL3.5-8B and LLaVA-OV-2-8B. Moreover, unlike Goal-Directed Social Reasoning, probe accuracy does not consistently exceed the corresponding single-video behavioral accuracy.

4.2 DOES LINEAR DECODABILITY PERSIST WHEN THE MODEL IS WRONG?

We next ask whether the judgment remains linearly decodable on examples that the model itself classifies incorrectly. We fix the *Last Query Token* representation at the final layer across all models and both reasoning domains. Using the held-out predictions from the same cross-validated probes, we partition videos according to whether the model's own single-video prediction is correct or incorrect. No separate probes are trained on these subsets. Table 2 reports the resulting accuracies.

For Goal-Directed Social Reasoning, Help/Hinder remains strongly linearly decodable even when the model is wrong: probe accuracy on the model-wrong subset ranges from 75.0% to 83.7%, compared with 79.6–90.3% when the model is correct, and for several models the gap is small. However, a successful probe establishes only that this distinction is linearly recoverable from the representation, not that the model relies on it when producing its answer.

Belief-Conditioned Rationality shows a different pattern. For the three Qwen models, probe accuracy falls from 70.5–81.8% when the model is correct to 49.0–58.3% when it is wrong. InternVL3.5-8B shows a similar drop from 64.6% to 50.3%, while LLaVA-OV-2-8B remains close to chance in both subsets. Taken together, the relationship between internal linear decodability and behavioral errors depends on the reasoning domain. For Goal-Directed Social Reasoning, the distinction can remain strongly linearly decodable even when the model produces the wrong answer, whereas for Belief-Conditioned Rationality, model errors largely coincide with a loss of linear decodability.

5 CONCLUSION

We present SMORE, to our knowledge the first benchmark to jointly evaluate belief-conditioned rationality, goal-directed social reasoning, and moral reasoning through controlled minimal-pair videos. Evaluating humans and state-of-the-art MLLMs, we find a substantial gap: averaged over all 427 pairs, humans reach 96% Joint-Pair Accuracy while the best model, Gemini-3.5-Flash, reaches only 63.2%. Because the two videos of each minimal pair differ in a single judgment-relevant variable, SMORE probes whether model judgments track relevant behavioral differences under closely matched conditions. When each video is judged on its own, models often give the same answer to both members of a pair. Yet when both videos are shown together, they can often identify which one exhibits the target property, suggesting that they detect the behavioral difference but struggle to judge each video independently. Models also show a consistent asymmetry: they are much more accurate when the correct judgment is favourable to the agent than when it is unfavourable. We complement these behavioral results with linear probing and find that, for goal-directed social reasoning, the judgment-relevant distinction can remain strongly linearly decodable from internal representations even when the model produces the wrong answer. We hope SMORE provides a useful benchmark for advancing MLLMs' ability to reason about the subtle differences in human behavior that determine social, moral, and rational judgments.

Ethics Statement

Our study includes human participants who evaluate the benchmark videos through a series of behavioral judgment quizzes. Participants were compensated at a rate above the applicable local minimum wage. The benchmark consists of synthetically generated or virtually rendered videos and does not contain personally identifiable information from the study participants. Human responses are reported only in aggregate. Additional details of the human evaluation procedure are provided in Appendix A.4.

Reproducibility Statement

Details required to reproduce benchmark construction are provided in Section 2.1 and Appendix A.6, while details required to reproduce the human and model evaluations are provided in Section 2.2 and Appendices A.4 and A.5.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

X. An et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

S. Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

X. Fan, X. Zhou, C. Jin, K. Nottingham, H. Zhu, and M. Sap. SoMi-ToM: Evaluating multi-perspective theory of mind in embodied social interactions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

K. Gandhi, G. Stojnic, B. M. Lake, and M. R. Dillon. Baby intuitions benchmark (BIB): Discerning the goals, preferences, and actions of others. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9963–9976, 2021.

K. Gandhi, J.-P. Fränken, T. Gerstenberg, and N. D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems*, 2023.

G. Gergely and G. Csibra. Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292, 2003.

Google DeepMind. Gemini 3.1 pro: Model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf`, 2026a.

Google DeepMind. Gemini 3.5 flash: Model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf`, 2026b.

Google DeepMind. Gemini omni flash: Model card. `https://deepmind.google/models/model-cards/gemini-omni-flash/`, 2026c.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019. Association for Computational Linguistics, 2024.

S. Guglielmo. Moral judgment as information processing: An integrative review. *Frontiers in Psychology*, 6:1637, 2015.

J. K. Hamlin, K. Wynn, and P. Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169): 557–559, 2007.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019.

Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. VIVA: A benchmark for vision-grounded decision-making with human values. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2294–2311. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.137.

Z. Huang, Z. Shi, H. Luo, H. Tang, M. Liu, M. Fang, and L. Chen. SocialReasonBench: A video-qa benchmark for social reasoning with counterfactual narrative videos. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025.

C. Jin, Y. Wu, J. Cao, J. Xiang, Y.-L. Kuo, Z. Hu, T. Ullman, A. Torralba, J. B. Tenenbaum, and T. Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085. Association for Computational Linguistics, 2026.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. SIV-Bench: A video benchmark for social interaction understanding and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

B. Krojer, M. Komeili, C. Ross, Q. Garrido, K. Sinha, N. Ballas, and M. Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Xiao Lin, Zhining Liu, Ze Yang, Gaotang Li, Ruizhong Qiu, Shuke Wang, Hui Liu, Haotian Li, Sumit Keswani, Vishwa Pardeshi, Huijun Zhao, Wei Fan, and Hanghang Tong. MORALISE: A structured benchmark for moral alignment in visual language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou. TempCompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

L. Niu, J. Li, X. Yu, X. Dong, S. Wang, R. Feng, B. Wu, P. Wei, Y. Wang, and L. Fan. Read the room: Video social reasoning with mental-physical causal chains. In *The Fourteenth International Conference on Learning Representations*, 2026.

Keane Ong, Sabri Boughorbel, Luwei Xiao, Chanakya Ekbote, Wei Dai, Ao Qu, Jingyao Wu, Rui Mao, Ehsan Hoque, Erik Cambria, Gianmarco Mengaldo, and Paul Pu Liang. OmniSapiens: A foundation model for social behavior processing via heterogeneity-aware relative policy optimization. In *Forty-third International Conference on Machine Learning*, 2026a.

Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, and Paul Pu Liang. Human behavior atlas: Benchmarking unified psychological and social behavior understanding. In *The Fourteenth International Conference on Learning Representations*, 2026b.

OpenAI. GPT-5.6 system card. `https://deploymentsafety.openai.com/gpt-5-6`, 2026.

N. Prakash, N. Shapira, A. S. Sharma, C. Riedl, Y. Belinkov, T. R. Shaham, D. Bau, and A. Geiger. Language models use lookbacks to track beliefs. In *The Fourteenth International Conference on Learning Representations*, 2026.

X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026a.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model. `https://qwen.ai/blog?id=qwen3.6-27b`, 2026b.

M. T. Rizzo and M. Killen. Morality, intentionality, and intergroup attitudes. *Behaviour*, 151(2–3): 337–359, 2014.

H. Shi, S. Ye, X. Fang, C. Jin, L. Isik, Y.-L. Kuo, and T. Shu. MuMA-ToM: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. B. Tenenbaum, and T. Ullman. AGENT: A benchmark for core psychological reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9614–9625, 2021.

T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

W. Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

S. F. White, H. Zhao, K. K. Leong, K. M. Smetana, and R. J. R. Blair. Neural correlates of conventional and harm/welfare-based moral decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, 17:1114–1128, 2017.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

B. Yan, J. Zhang, Z. Chen, S. Shan, and X. Chen. M^3oralBench: A multimodal moral benchmark for lvlms. *arXiv preprint arXiv:2412.20718*, 2024.

B. Yang et al. Kwai Keye-VL 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025b.

J. Zhang, M. Cai, and Y. J. Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

A Appendix

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

We use LLMs in the following aspects: (i) We utilized LLMs in the dataset construction process; (ii) We evaluated task performance on several LLMs; (iii) We used LLMs to assist us with paper writing slightly.

A.2 Limitations

SMORE uses virtually rendered and generated videos rather than real footage. This provides the control that minimal pairs require, but performance may not transfer directly to naturalistic videos. The two generation routes also offer different guarantees. VirtualHome pairs are controlled by construction, whereas generated Gemini Omni pairs rely on post-generation verification, so undetected residual visual differences may remain. The Moral subset is also smaller, so its subdomain estimates carry wider uncertainty, and because these videos are generated by a Gemini model, we cannot rule out model-family effects when evaluating Gemini models. Finally, our reasoning-trace analysis is behavioral: it characterizes what models write rather than the computation underlying their answers; proprietary models expose only reasoning summaries, and traces are annotated by an LLM.

A.3 Related Work

Recent benchmarks evaluate increasingly rich forms of mental-state and social reasoning in multimodal models. Video benchmarks such as SoMi-ToM (Fan et al., 2025), MoMentS (Villa-Cueva et al., 2025), SIV-Bench (Kong et al., 2026), and R^3-Bench (Niu et al., 2026) evaluate mental-state and social understanding by posing questions about individual videos. MMToM-QA (Jin et al., 2024) evaluates belief and goal inference from multimodal observations of human behavior, while MuMA-ToM (Shi et al., 2025) extends this setting to multi-agent interactions, including beliefs, social goals such as whether agents help or hinder one another, and beliefs about others' goals. These tasks primarily require *inverse* reasoning from observed behavior to the latent beliefs or goals that could have generated it, typically under an assumption of approximately rational action. In contrast, for belief-conditioned rationality, SMORE directly evaluates the corresponding *forward* judgment (Gergely & Csibra, 2003): given an agent's stated belief and goal, is its observed action rational? Earlier benchmarks such as AGENT (Shu et al., 2021) and BIB (Gandhi et al., 2021) probe related notions of instrumental rationality, including whether agents act efficiently toward their goals in simple synthetic environments. However, they do not test how the rationality of the same action changes with the agent's belief, and their expected–unexpected contrasts are primarily evaluated through violation-of-expectation rather than individual assessment.

Multimodal moral reasoning has likewise received increasing attention. VIVA (Hu et al., 2024) evaluates vision-grounded decision-making with human values, M^3oralBench (Yan et al., 2024) evaluates moral judgment and classification from generated image–text scenarios, and MORALISE (Lin et al., 2026) uses real-world image–text pairs to evaluate moral judgment and norm attribution. These benchmarks broaden moral evaluation beyond text, but remain based on static visual inputs. SMORE instead evaluates moral judgments from videos in which temporally unfolding behavior, such as an agent's intention or response to the consequences of an action, determines the judgment.

Controlled contrasts have been explored separately in both Theory-of-Mind and video understanding. In language, BigToM (Gandhi et al., 2023) constructs controlled causal variations of text-based social scenarios to test belief and action inference, while CausalToM (Prakash et al., 2026) constructs counterfactual text pairs for controlled analysis of belief tracking. These evaluations, however, remain text-only. In video, SocialReasonBench (Huang et al., 2026) uses the branching storylines of an interactive narrative game to ground counterfactual social questions, but its alternative branches can differ in many respects at once rather than in a single controlled factor. For other aspects of video understanding, TempCompass (Liu et al., 2024) constructs conflicting videos that retain similar static content while changing temporal properties, Vinoground (Zhang et al., 2024) uses temporal counterfactual video contrasts, and MVP (Krojer et al., 2025) explicitly evaluates minimally different video pairs for physical understanding. These benchmarks demonstrate the value of controlled video contrasts, but focus primarily on temporal and physical understanding rather than social-cognitive judgments. Winoground (Thrush et al., 2022) and MVP (Krojer et al., 2025) further score models at

the level of the contrast, crediting a model only when it answers every member correctly; our Joint-Pair Accuracy follows this practice. SMORE brings these directions together by using minimally different videos to evaluate belief-conditioned rationality, goal-directed social reasoning, and moral reasoning under controlled behavioral changes.

A.4 Human Evaluation

Participants and coverage. Human evaluation covers the full benchmark: all 1,281 questions across the 427 minimal pairs, comprising 854 single-video questions and 427 paired-video questions. Each question was answered independently by three annotators drawn from a pool of 65 annotators.

Annotation protocol. Annotators were evaluated under the same protocol as the models and were shown the same video(s), question text, and answer options. In the single-video setting, annotators viewed one video and answered a two-option multiple-choice question about the target judgment, for example whether an action was rational or irrational, or whether an agent helped or hindered another. In the paired-video setting, annotators viewed both videos of a minimal pair and answered a two-option multiple-choice question asking which of the two videos exhibited the target property.

Aggregation and reliability. For each question, we take the majority vote answers of the three annotations as the human answer. Human Paired, Single-video, and Joint-Pair Accuracy are then computed from these majority-vote. We measure inter-annotator agreement using Fleiss' κ, computed separately for the single-video and paired-video settings. For Belief-Conditioned Rationality, Fleiss' κ is 0.82 in the single-video setting and 0.75 in the paired-video setting; for Goal-Directed Social Reasoning, it is 0.85 in the single-video setting and 0.78 in the paired-video setting. Within Moral Reasoning, Fleiss' κ is 0.78 in the single-video setting and 1.00 in the paired-video setting for Intentionality; 0.95 in the single-video setting and 1.00 in the paired-video setting for Responsibility Taking; and 0.81 in the single-video setting and 0.89 in the paired-video setting for Moral/Conventional Transgressions.

A.5 Model Evaluation Details

For all model evaluations, each video is represented by 40 frames uniformly sampled over its full duration. In the single-video setting, the model is provided with the sampled frames, the corresponding question, and two answer options. In the paired-video setting, each video is represented by 40 uniformly sampled frames, and both videos are presented jointly with the corresponding two-option question, with the assignment of the two videos to Video A and Video B randomized for each pair. The same question and answer-option wording is used across models.

We use greedy decoding for all model evaluations to ensure deterministic inference and reproducibility. Models are instructed to answer with one of the provided answer options, and the selected option is used directly to compute accuracy.

A.6 Data Generation

Videos are produced by two generation routes, matched to the control requirements of each reasoning domain. Belief-conditioned rationality and goal-directed social reasoning require exact control over object identity, spatial layout, navigation cost, and the timing of one agent's action relative to another's, so these are generated in a household simulator where each of these factors is an explicit parameter (350 pairs, 700 videos). Moral reasoning scenarios turn on intentionality, responsibility, and the violation of social norms, which fall outside the fixed action vocabulary and object affordances of a simulator, so these are generated with a text-and-image-conditioned video model (77 pairs, 154 videos). In both routes, everything that is not the manipulated variable is fixed at generation time rather than just being matched after the fact. Section A.6.1 describes the simulator route and Section A.6.2 the generative route.

A.6.1 Procedurally generated videos (VirtualHome)

Rationality and goal-directed social reasoning videos are generated in VirtualHome (Puig et al., 2018). Scenarios are specified as scenario templates, filled by a deterministic seeded sampler, and executed

Action	Arity	Precondition
[walk]	1	target reachable from the actor's position
[grab]	1	target grabbable; actor adjacent; hands free
[open]	1	target a container, currently closed
[close]	1	target a container, currently open
[putin]	2	actor holding the first target; second target an open container
[putback]	2	actor holding the first target; second target a surface
[switchon]	1	target a device, currently off
[switchoff]	1	target a device, currently on

Table 3: Action vocabulary used by the scripted runners. Preconditions are checked against the environment graph when an instance is sampled and again by the simulator at execution time.

by scripted runners: every action performed by every agent is emitted by the template, so the behavior in each episode is fully determined by the template and the sampled configuration.

Scenario templates. A template specifies the role structure of a scenario: which agent holds the goal and which agent intervenes, together with the object classes each role requires, the ordering of the action sequence, and the affordance constraints a sampled instance must satisfy. Templates are kept small, so that the manipulated variable occupies a single position in the sequence. Each reasoning condition reported in the main text corresponds to one template family: true and false belief and the safe and risky path conditions for rationality, and FIND, OPEN/CLOSE, and ON/OFF for goal-directed social reasoning.

Environments and sampling. VirtualHome provides a set of fixed indoor apartments, each represented as an environment graph whose nodes are objects with a semantic class, an instance identifier, and a state, and whose edges encode spatial relationships. Objects within an environment can be added, removed, or repositioned programmatically, so a single apartment supports many distinct scenes.

For each episode the sampler draws an environment and a room from the 15 apartments we use, reads the environment graph for that room, and selects from it the two agent agents, the goal object instance, the container or device instance the scenario acts on, and any auxiliary targets such as the intervening agent's walk destination. Selection is constrained by the affordances recorded in the graph: a [putin] destination must be a container with interior capacity, a [putback] destination must be a surface, items to be moved must be grabbable and reachable, and the goal object must not already occupy its destination. Instances violating a constraint are rejected and redrawn. Sampling is seeded and the seed is recorded with the episode, so any episode can be reproduced exactly; individual selections can also be fixed rather than sampled, which we use to hold a particular apartment or object constant while varying the rest.

Agent appearance is recorded in natural language alongside the agent identifier (for example, *the agent in the light blue suit*), giving questions a referring expression that is unambiguous from the sampled viewpoint and identical across the two videos of a pair.

Action representation. Instantiated templates are emitted as scripts. Each line names an actor, an atomic action, and its targets, where a target is an object class paired with the instance identifier of that object in the sampled environment graph:

```
<char1> [putin] <creamybuns> (120) <dishwasher> (165)
```

The action vocabulary is listed in Table 3. Two lines joined by | are executed concurrently within a single timestep, which places one agent's intervention at a precise point in the other agent's activity. Instance identifiers are identical across the two videos of a pair, so a given object is the same instance in the same location in both.

Minimal-pair construction. The two videos of a pair are rendered from a single sampled configuration. The apartment, room, seed, agents and their appearance, object instances, the goal held by the acting agent, and the camera pose are shared between them. What is manipulated is one judgment-relevant element, realized by one of three operators:

1. **Action substitution.** One action of the intervening agent on a shared target is replaced by its counterpart: `[open]` for `[close]`, `[switchon]` for `[switchoff]`, or a destination that brings the goal object nearer to rather than farther from the requesting agent.
2. **Timed scene-state change.** The intervening agent's action is placed concurrently with a specific action of the acting agent, so that the state of the scene changes at a controlled point in that agent's activity for instance, a container being closed in the same timestep in which the other agent picks up the object it intends to place there.
3. **Route manipulation.** The route available to the acting agent is altered by invisible colliders while start and goal positions are held fixed, so that the same destination is reached by paths of different cost.

Minimality is a property of the rendered videos. The two videos show the same agents, in the same apartment and room, from the same viewpoint, acting on the same objects toward the same goal, and differ visibly in the manipulated element alone. The scripts are given in full so that what each agent does, and does not do, is explicit.

Worked example: OPEN/CLOSE. Figure 6 gives both videos of one OPEN/CLOSE pair, rendered from the same configuration: apartment 2, kitchen, seed 91, the same two agents, the same camera pose, the `creamybuns` (120) and the `dishwasher` (165). The acting agent `char0` has the goal of placing the creamybuns in the dishwasher. In the helping video, `char1` opens the dishwasher and then moves away, and `char0` completes the placement. In the hindering video, `char0` opens the dishwasher himself, and `char1` closes it in the same timestep in which `char0` picks up the creamybuns; `char0` returns to the dishwasher and the placement does not occur.

Helping

```
<char1> [walk]  <dishwasher> (165)
<char1> [open]  <dishwasher> (165)
<char0> [walk]  <creamybuns>  (120)
<char0> [grab]  <creamybuns>  (120)
<char1> [walk]  <coffeemaker> (169)
<char0> [walk]  <dishwasher> (165)
<char0> [putin] <creamybuns>  (120) <dishwasher> (165)
<char0> [close] <dishwasher> (165)
```

Hindering

```
<char0> [walk]  <dishwasher> (165)
<char0> [open]  <dishwasher> (165)
<char0> [walk]  <creamybuns>  (120)
<char1> [close] <dishwasher> (165)  |  <char0> [grab] <creamybuns> (120)
<char0> [walk]  <dishwasher> (165)
```

Figure 6: Both videos of one OPEN/CLOSE pair. Agents, apartment, room, seed, camera pose, object instances, and goal are shared; the manipulated element, marked in red, is the intervening agent's action on the container. The placing action is absent from the hindering video because the container is closed before the acting agent reaches it.

Rendering and verification. Videos are rendered at 1280×720 and 10 frames per second, with a mean length of 12.3 seconds. Within a pair the camera strategy, index, and pose are identical. The simulator reports a per-agent status for each executed script. An episode in which execution does not succeed in case a precondition is proved unsatisfiable in the sampled scene, or navigation failed is then retried and otherwise resampled, and only episodes that execute successfully are rendered. Every rendered video was then verified by the authors: that the manipulated element is visible from the sampled viewpoint and not occluded at the moment it occurs, that the goal action visibly completes

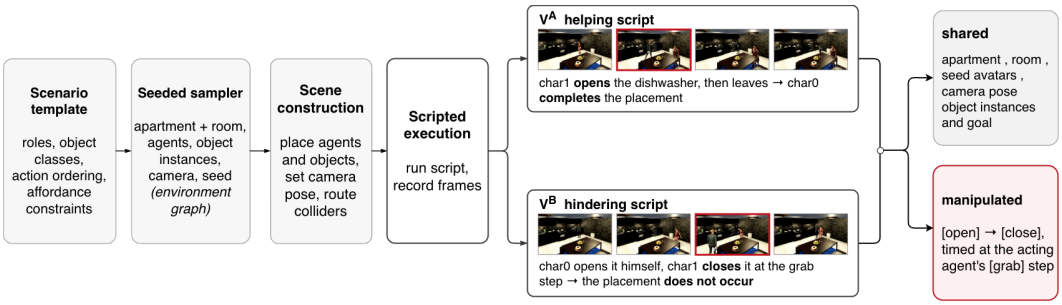

Figure 7: Generation pipeline. Procedural route: a scenario template is filled by a seeded sampler against the environment graph of the selected apartment, the scene is constructed, and the scripts are executed. Both videos of a pair come from one configuration, so the listed factors are shared and only the manipulated element differs.

or visibly fails as the condition requires, and that the two videos of a pair differ in the manipulated element and its consequences alone. Pairs failing any check were regenerated under a new.

A.6.2 Generatively produced videos (Gemini Omni)

Moral reasoning videos are generated with Gemini Omni (Google DeepMind, 2026c). The three reasoning domains: intentionality, responsibility taking, and moral versus conventional transgressions require behavior that has no counterpart in a simulator's action vocabulary: an action performed deliberately rather than by accident, an agent owning or deflecting a negative outcome it caused, a violation of a norm as against a convention. Generation is controlled along two axes: visual conditioning, which holds the actors and the location fixed, and a prompt scaffold, which holds the scene specification fixed and confines the manipulation to a single sentence.

Visual conditioning. Each pair is conditioned on three reference images: one per actor and one of the location. The actor references are frontal, neutral-expression, plain-background portraits, and the location reference is a wide view of the room in which the interaction takes place. The prompt binds each reference to a named role and instructs the model to treat the location image as the exact setting, preserving its layout, furnishings, lighting, and visual style. The two videos of a pair are conditioned on the identical set of reference images, so actor identity and appearance and the room itself are held fixed across the manipulation rather than re-specified in words.

All reference identities are synthetic, no real person's likeness is used anywhere in the benchmark.

Prompt scaffold. Prompts are assembled from a fixed scaffold whose slots are listed in Table 4. The scaffold fixes camera treatment (a static, locked-off wide shot in one continuous take), framing (both actors visible in frame throughout), cast closure (the two named actors are the only people present, and none enter in the background), the lighting and documentary look, the audio specification (diegetic room tone and incidental sound, no dialogue), and a list of exclusions covering on-screen text, subtitles, captions, slow motion, music, and voiceover. The scaffold then states the shared setup: the actions leading up to the judgment-relevant moment, identical in both members of the pair.

Fixing these properties in the prompt is what makes the resulting videos comparable. The camera and framing constraints keep the manipulated event inside the frame and remove shot composition as a variable, cast closure prevents the model from populating the background differently across two generations. The no-dialogue constraint keeps the judgment grounded in what the agents do, so that the manipulated variable is an action and not a verbal statement of intent.

Minimal-pair construction. The two prompts of a pair are identical strings up to their final sentence, and are paired with the identical reference images. The final sentence specifies the judgment-relevant action and is the only manipulated element: in responsibility taking, whether the agent who caused the accident remedies it or leaves; in intentionality, whether the agent performs an action that is designed to appear deliberate or one that appears accidental; in moral versus conventional transgressions, whether the violated rule concerns harm, rights, or property, or a context-dependent social custom.

Slot	Status	Content
Reference binding	fixed	each actor image bound to a name; location image as the exact setting
Camera	fixed	static locked-off wide shot, one continuous take, no cuts
Framing	fixed	both actors fully visible at all times
Cast closure	fixed	the two actors only; no background people
Look	fixed	warm even natural light; documentary style
Audio	fixed	diegetic room tone and incidental sound; no dialogue
Exclusions	fixed	no text, subtitles, captions, slow motion, music, voiceover
Setup	fixed	the actions preceding the judgment-relevant moment
Outcome	**varied**	the single sentence specifying the judgment-relevant action

Table 4: Prompt scaffold for the generative route. Every slot is identical across the two videos of a pair except the final one.

Multiple candidate videos were generated for each prompt. For each scenario, the two videos forming the minimal pair were selected jointly by the authors to maximize visual consistency across the pair while clearly realizing the intended behavioral manipulation. Preference was given to pairs without noticeable generation artifacts; where minor artifacts were present, a video was retained only if they were unrelated to the judgment-relevant behavior and could not affect the target judgment. Each selected pair was then verified by the authors to ensure that the intended manipulation was clearly realized and that no other salient difference between the videos could influence the judgment.

Worked example: responsibility taking. Figure 8 reproduces a complete prompt pair. The shared body runs to the collision; the manipulated sentence determines whether the agent who caused it takes responsibility. Figure 9 shows the reference images and keyframes from the resulting videos.

Acceptance and verification. Generated clips are 1280×720 at 24 frames per second , with length of 6 to 8 seconds. A clip is accepted only if it satisfies the properties the scaffold specifies: the actors match their reference images and remain recognizable throughout, the location matches the reference, the shot is a single continuous take with a static camera, both actors stay in frame, no additional people appear, no dialogue or editorial overlay is present, the setup event is clearly visible, and the outcome is the one the prompt specified. Four clips were generated for each scenario, from which the most suitable clip was selected based on the acceptance criteria. Both members of each pair were then verified together to ensure that they differed in the intended manipulated outcome while remaining otherwise visually consistent.

A.6.3 Representation-level validation of minimal-pair control

We additionally test whether the constructed minimal pairs are visually matched in the representation space of a video MLLM. This analysis is intended as a validation of the benchmark construction rather than as a measure of model reasoning: if two videos form a controlled minimal pair, they should be substantially more similar to one another than to videos belonging to different pairs.

We perform this analysis using the vision encoder of Qwen3.6-27B. Each video is uniformly sampled to 40 frames and encoded using the final vision-encoder representation. The vision encoder internally groups consecutive frames in pairs, yielding $T = 20$ temporally ordered groups. Within each temporal group, we spatially average the corresponding patch-token features. Thus, each video V is represented as

$$X(V) = [x_1(V), \ldots, x_T(V)] \in \mathbb{R}^{T \times H},$$

where $T = 20$ and the vision-encoder hidden dimension is $H = 1152$. Importantly, the spatial patch dimension is averaged out, while the temporal dimension is retained.

Shared body (both videos)

Using the provided image of Maya as Maya, the provided image of Marcus as Marcus, and the provided gym image as the exact location -- preserve the gym's layout, furniture, lighting, and overall visual style as closely as possible. Static, locked-off wide shot, one continuous take with no cuts. Keep Maya and Marcus fully visible in frame at all times. Maya and Marcus are the only two people present; no other people enter or appear in the background. Warm, even natural lighting; realistic fly-on-the-wall documentary look. No on-screen text, subtitles, captions, slow motion, music, voiceover, or spoken dialogue. Diegetic audio only: quiet gym room tone with faint equipment clinks and a bump and items clattering across the floor; no dialogue -- neither Maya nor Marcus speaks at any point. In a quiet gym, Marcus walks along carrying a stack of towels and a water bottle. Maya, hurrying around a corner, accidentally collides with him and he drops everything on the floor.

Outcome sentence, responsible

Maya stops and gathers up the towels and water bottle, handing them back to Marcus.

Outcome sentence, not responsible

Maya walks away out of the frame, leaving Marcus among the dropped towels.

Figure 8: A complete prompt pair for the responsibility-taking component. The body above the rule is shared verbatim; the two videos differ only in the sentence below it, shown in red. Both prompts are conditioned on the same three reference images.

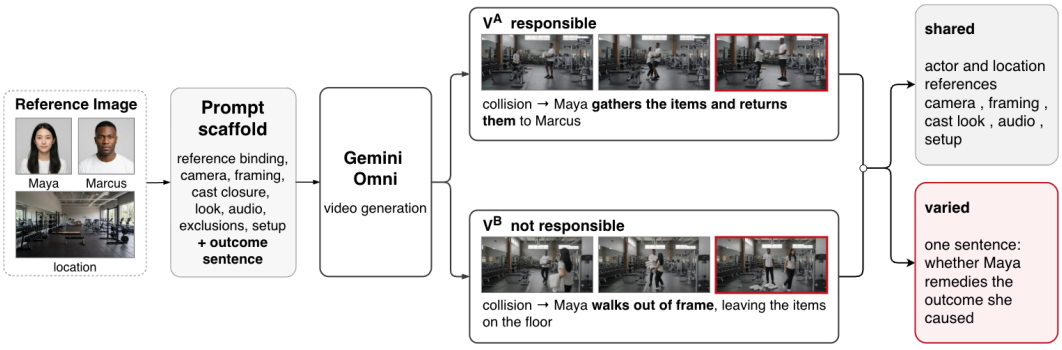

Figure 9: Generative route. The same three reference images: one per actor and one of the location and the same prompt body condition both generations; only the outcome sentence differs. Keyframes are drawn from the accepted pair for the responsibility-taking example of Figure 8.

Before computing distances, we standardize each of the H representation dimensions independently across all videos and temporal positions. Let $x_{n,t,h}$ denote dimension h of temporal group t for video n. For each hidden dimension h, we compute

$$\mu_h = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} x_{n,t,h}, \qquad \sigma_h = \sqrt{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left(x_{n,t,h}-\mu_h\right)^2},$$

and obtain the standardized representation

$$\hat{x}_{n,t,h} = \frac{x_{n,t,h}-\mu_h}{\sigma_h}.$$

We then ℓ_2-normalize each temporal feature vector,

$$z_t(V) = \frac{\hat{x}_t(V)}{\|\hat{x}_t(V)\|_2}.$$

Here, $t \in \{1, \ldots, T\}$ denotes the temporal-group index. Since every video is uniformly sampled to the same number of frames, corresponding values of t represent corresponding relative positions within the two clips. We first define the cosine distance at temporal position t as

$$d(X, Y; t) = 1 - \cos(z_t(X), z_t(Y)),$$

where only features at the same temporal-group index are compared. We then define the aligned distance between two videos by averaging these distances over temporal positions:

$$d(X, Y) = \frac{1}{T} \sum_{t=1}^{T} d(X, Y; t) = \frac{1}{T} \sum_{t=1}^{T} [1 - \cos(z_t(X), z_t(Y))].$$

Thus, temporal correspondence is preserved when computing similarity, while $d(X, Y)$ provides a single distance between the two videos.

We treat every minimal pair as an unordered set

$$P_i = \{V_i^{(1)}, V_i^{(2)}\}, \qquad i = 1, \ldots, M.$$

The average within-pair distance is

$$D_{\text{within}} = \frac{1}{M} \sum_{i=1}^{M} d\Big(V_i^{(1)}, V_i^{(2)}\Big).$$

To construct an unrelated-video baseline, we compare every pair of distinct minimal pairs P_i and P_j. For a pair of minimal pairs, we average over all four possible cross-pair video comparisons:

$$d_{\text{cross}}(P_i, P_j) = \frac{1}{4} \sum_{a=1}^{2} \sum_{b=1}^{2} d\Big(V_i^{(a)}, V_j^{(b)}\Big).$$

The dataset-level cross-pair distance is then

$$D_{\text{cross}} = \frac{1}{\binom{M}{2}} \sum_{i<j} d_{\text{cross}}(P_i, P_j).$$

Finally, we summarize the relative within-pair similarity as

$$R = \frac{D_{\text{within}}}{D_{\text{cross}}}.$$

Values of $R < 1$ indicate that videos belonging to the same minimal pair are, on average, closer in visual representation space than videos belonging to different pairs.

We estimate uncertainty by bootstrapping at the level of minimal pairs. For each of 10,000 bootstrap replicates, we resample the M minimal pairs with replacement and recompute D_{within}, D_{cross}, and R. We report percentile 95% confidence intervals for R.

Across all five tasks, within-pair distances are substantially smaller than cross-pair distances, and every 95% confidence interval for R lies well below 1 (Table 5). Thus, videos belonging to the same minimal pair are consistently much closer in the model's visual representation space than videos drawn from different pairs, providing representation-level evidence that the constructed pairs preserve substantially more of their underlying visual scenario than unrelated videos.

A.6.4 Human validation of minimal-pair control

To complement the construction details and metric-based analysis above, we ran a human study on a representative subset of 13 video pairs that covers all ten subclasses and includes both VirtualHome

Task	$\mathbf{D}_{\text{within}}$	$\mathbf{D}_{\text{cross}}$	R	95% CI
Goal-Directed Social Reasoning	0.154	0.993	0.155	[0.133, 0.178]
Responsibility Taking	0.198	1.017	0.194	[0.153, 0.244]
Intentionality	0.204	1.011	0.202	[0.155, 0.254]
Moral/Conventional	0.309	1.024	0.302	[0.237, 0.367]
Belief-Conditioned Rationality	0.412	0.822	0.501	[0.481, 0.523]

Table 5: **Representation-level validation of minimal-pair similarity.** Average within-pair and cross-pair cosine distances at the final vision-encoder representation of Qwen3.6-27B. $R = D_{\text{within}}/D_{\text{cross}}$; lower values indicate that videos belonging to the same minimal pair are more similar than videos from different pairs. Confidence intervals are obtained by bootstrapping minimal pairs.

and Omni-generated videos. For each pair, annotators first described the key difference between the two videos without seeing the benchmark question. They then saw the question and judged whether the pair forms a minimal pair under our definition, whether they noticed any visual or generation artifacts, and whether any artifact could reveal the answer. In nearly all cases, annotators described the intended behavioral contrast rather than low-level visual differences, and they rated most pairs as minimal or mostly minimal. Minor artifacts were occasionally reported, but annotators almost never judged that these artifacts alone could reveal the answer. We treat this study more as a sanity check rather than a full validation of minimal-pair control.

A.7 Qualitative Model Example

Figure 10 shows a representative example from Gemini 3.1 Pro on a Goal-Directed Social Reasoning minimal pair from. The question asked for both videos is:

> The video shows the agent in the red patterned dress and the agent in the grey sweatshirt. The goal of the agent in the red patterned dress is to get the book.
>
> In this video, which of the following is true?
>
> 1. The agent in the grey sweatshirt helps the agent in the red patterned dress achieve the goal.
>
> 2. The agent in the grey sweatshirt hinders the agent in the red patterned dress from achieving the goal.

Video A (helper). Gemini 3.1 Pro answers **1**; the ground-truth answer is **1**, so the prediction is correct.

> **Analyzing the Agent Interaction**
>
> Okay, here's the breakdown of this scenario. I need to figure out if the guy in the grey sweatshirt is helping or hindering the woman in the red dress.
>
> First, let's identify the players. We have a woman near the bed, the "red patterned dress agent," and a guy moving around, the "grey sweatshirt agent." Now, what's the objective? Ah, she wants the book. She explicitly asks for it.
>
> Next, I observe what the guy in grey does. He heads for the shelves, looks at them, grabs something – looks like a book or folder – and then turns around. He's **carrying** this object and walking toward her, implying he is bringing her the item.
>
> Evaluating the interaction. She **asked** for the book; he went and **got** it. He is **bringing** it to her. So, it's pretty clear: he's helping.
>
> Therefore, since the agent in the grey sweatshirt is helping the agent in the red patterned dress achieve her goal, the answer is 1.

Video B (hinder). Gemini 3.1 Pro answers **1**; the ground-truth answer is **2**, so the prediction is incorrect.

> **Analyzing the Interaction: Does the Grey Sweatshirt Help?**
>
> Okay, let's break this down. My objective is to determine if the agent in the grey sweatshirt is assisting or obstructing the agent in the red patterned dress. First, I establish the key players: the woman in the red dress and the man in the grey sweatshirt. The woman has a



Figure 10: Qualitative example from the `hh_book` minimal pair. Gemini 3.1 Pro predicts **help** for both videos in the single-video setting, correctly on Video A and incorrectly on Video B.

> clearly defined goal – she wants a book. She makes that explicitly clear at the timestamp, "Can you please fetch me the book?"
>
> Now, I need to analyze the man's actions. He turns and heads straight for the shelves, specifically the white shelving unit. He seems to examine a book or binder and picks something up. He then walks off with the book in his hand. He hasn't yet returned it to her, but the immediate action of fetching the book aligns with her request.
>
> Essentially, the woman **asks** for the book, and the man **fetches** it. The man is actively working toward her objective. Therefore, he's providing assistance. Option 1 confirms that he helps. Option 2 would state he hinders her, which is just wrong.

As shown in the reasoning traces, Gemini 3.1 Pro correctly interprets Video A: the agent retrieves the requested book and brings it toward the other agent, and the model appropriately reasons that this constitutes helping. In Video B, however, the agent retrieves the book and then carries it away and places it at a distant location, thereby hindering the other agent's goal. Despite this difference, the model still reasons that the initial act of fetching the book constitutes helping and predicts the same label as in Video A.

A.8 Additional Results and Analyses

This section reports additional observations drawn from the complete results in Table 6, which lists Paired (P), Single-video (S), and Joint-Pair Accuracy (J) for every evaluated model and subdomain, and from the evaluation of human-behavior-specialized MLLMs in Table 8. Unless otherwise noted, comparisons use J, which credits a minimal pair only when both videos are judged correctly (chance: 25%).

A.8.1 Models struggle most on Believed Safe, Believed Risky, and Find

The largest failures in Joint-Pair Accuracy are concentrated in three subdomains: Believed Safe, Believed Risky, and Find. Across the full model suite (Table 6), many models fall *below* the 25% chance level; 8 of 13 models on Believed Safe, 9 of 13 on Believed Risky, and 11 of 13 on Find. Only Gemini-3.1-Pro and Gemini-3.5-Flash remain above chance across all three subdomains. These subdomains share an important property in the benchmark design. In the Believed Safe and Believed Risky subdomains, whether an action is rational depends on the path trajectory an agent takes given its stated belief. A longer route can therefore be either rational or irrational depending on what the agent believes. Similarly, in Find, whether moving an object helps or hinders depends on where the agent moves it relative to the other agent's position. Thus, succeeding on these subdomains requires models to keep track of agents' trajectories and evaluate their actions relative to the relevant belief

Table 6: **Evaluation results on SMORE (%).** We report three metrics: Paired accuracy (P), Single-video accuracy (S), and Joint-Pair Accuracy (J), as defined in Section 2.2. Human scores use three independent annotations per question from a pool of 65 annotators. Values shown in red indicate Joint-Pair Accuracy scores below chance level.

	Rationality														
	True Belief			**False Belief**			**Believed Safe**			**Believed Risky**			**Overall**		
Model	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J
Human	**100.0**	**98.0**	**96.0**	**100.0**	**95.0**	**90.0**	**100.0**	**99.0**	**98.0**	**98.0**	**100.0**	**100.0**	**99.5**	**98.0**	**96.0**
Chance	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0
InternVL3.5-8B Inst	54.0	63.0	26.0	54.0	50.0	6.0	40.0	52.0	8.0	44.0	58.0	22.0	48.0	55.8	15.5
LLaVA-OV-2-8B Inst	52.0	61.0	30.0	56.0	50.0	18.0	62.0	57.0	14.0	36.0	46.0	10.0	51.5	53.5	18.0
Keye-VL-1.5-8B Think	56.0	74.0	56.0	66.0	76.0	54.0	58.0	49.0	14.0	50.0	42.0	12.0	57.5	60.3	34.0
Qwen3.5-9B Inst	92.0	60.0	22.0	34.0	53.0	12.0	56.0	50.0	6.0	58.0	51.0	8.0	60.0	53.5	12.0
Qwen3.5-9B Think	96.0	90.0	80.0	84.0	80.0	60.0	78.0	60.0	22.0	44.0	55.0	26.0	75.5	71.3	47.0
Qwen3.5-27B Inst	94.0	72.0	44.0	62.0	66.0	34.0	96.0	51.0	4.0	20.0	47.0	4.0	68.0	59.0	21.5
Qwen3.5-27B Think	98.0	94.0	88.0	88.0	80.0	60.0	92.0	59.0	28.0	72.0	57.0	20.0	87.5	72.5	49.0
Qwen3.6-27B Inst	94.0	91.0	84.0	66.0	79.0	64.0	78.0	55.0	14.0	46.0	47.0	0.0	71.0	68.0	40.5
Qwen3.6-27B Think	100.0	92.0	84.0	88.0	82.0	64.0	90.0	60.0	32.0	66.0	62.0	32.0	86.0	74.0	53.0
Qwen3-VL-32B Think	96.0	92.0	84.0	92.0	90.0	80.0	88.0	62.0	28.0	74.0	44.0	8.0	87.5	72.0	50.0
Gemini-3.1-Pro	96.0	93.0	86.0	98.0	94.0	88.0	74.0	67.0	42.0	58.0	69.0	50.0	81.5	80.8	66.5
Gemini-3.5-Flash	96.0	89.0	80.0	98.0	92.0	84.0	98.0	65.0	32.0	76.0	76.0	52.0	92.0	80.5	62.0
GPT5.6 Luna	96.0	83.0	66.0	94.0	77.0	54.0	96.0	58.0	16.0	78.0	58.0	24.0	91.0	69.0	40.0

	Goal-Directed Social Reasoning												**Moral Reasoning**								
	Find			**Open/Close**			**On/Off**			**Overall**			**Int.**			**Resp.**			**Moral/Conv.**		
Model	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J
Human	**100.0**	**96.6**	**93.2**	**97.9**	**98.9**	**97.9**	**100.0**	**98.9**	**97.7**	**99.3**	**98.0**	**96.0**	**100.0**	**98.0**	**96.0**	**100.0**	**98.1**	**96.3**	**100.0**	**98.0**	**96.0**
Chance	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0
InternVL3.5-8B Inst	49.2	56.8	15.3	74.5	43.6	4.3	47.7	50.0	4.5	56.7	50.7	8.7	60.0	58.0	20.0	88.9	87.0	74.1	48.0	50.0	0.0
LLaVA-OV-2-8B Inst	50.8	55.1	13.6	72.3	48.9	6.4	59.1	52.3	9.1	60.0	52.3	10.0	72.0	62.0	24.0	81.5	87.0	74.1	72.0	56.0	12.0
Keye-VL-1.5-8B Think	49.2	52.5	13.6	59.6	57.4	29.8	54.5	54.5	29.5	54.0	54.7	23.3	64.0	60.0	20.0	74.1	77.8	55.6	92.0	54.0	8.0
Qwen3.5-9B Inst	45.8	50.0	0.0	48.9	53.2	6.4	43.2	55.7	13.6	46.0	52.7	6.0	68.0	84.0	68.0	96.3	81.5	63.0	96.0	54.0	8.0
Qwen3.5-9B Think	64.4	51.7	3.4	80.9	61.7	38.3	68.2	65.9	36.4	70.7	59.0	24.0	84.0	76.0	52.0	96.3	90.7	81.5	100.0	72.0	44.0
Qwen3.5-27B Inst	55.9	51.7	3.4	68.1	69.1	44.7	50.0	73.9	52.3	58.0	63.7	30.7	100.0	78.0	56.0	100.0	92.6	85.2	96.0	66.0	32.0
Qwen3.5-27B Think	83.1	60.2	22.0	72.3	79.8	66.0	75.0	68.2	52.3	77.3	68.7	44.7	100.0	84.0	68.0	100.0	100.0	100.0	100.0	72.0	44.0
Qwen3.6-27B Inst	69.5	52.5	5.1	46.8	56.4	14.9	45.5	73.9	52.3	55.3	60.0	22.0	96.0	76.0	52.0	96.3	94.4	88.9	84.0	76.0	52.0
Qwen3.6-27B Think	88.1	61.9	23.7	80.9	80.9	68.1	75.0	76.1	54.5	82.0	72.0	46.7	96.0	76.0	52.0	96.3	98.1	96.3	100.0	82.0	64.0
Qwen3-VL-32B Think	54.2	50.8	5.1	70.2	73.4	48.9	56.8	78.4	56.8	60.0	66.0	34.0	96.0	80.0	60.0	88.9	90.7	81.5	96.0	58.0	16.0
Gemini-3.1-Pro	52.5	60.2	32.2	68.1	61.7	38.3	63.6	73.9	47.7	60.7	64.7	38.7	100.0	92.0	84.0	100.0	98.1	96.3	100.0	84.0	68.0
Gemini-3.5-Flash	78.0	72.9	47.5	55.3	72.3	57.4	77.3	75.0	50.0	70.7	73.3	51.3	96.0	96.0	92.0	100.0	96.3	92.6	100.0	92.0	84.0
GPT5.6 Luna	76.3	54.2	11.9	70.2	76.6	55.3	79.5	75.0	59.1	75.3	67.3	39.3	76.0	94.0	88.0	88.9	92.6	85.2	100.0	74.0	48.0

and goal. Reasoning traces support this interpretation: on FIND, incorrect traces of both 27B Qwen models treat picking up the requested object as evidence of helping, without following where it ultimately ends up (Appendix A.9). Humans perform near ceiling here, suggesting they reliably track these trajectories and use them to make the right judgments.

A.8.2 REASONING-ENABLED INFERENCE IMPROVES PAIR-LEVEL RELIABILITY

Matched Qwen Instruct and Think variants consistently benefit from reasoning-enabled inference. For Qwen3.6-27B, Joint-Pair Accuracy increases from 40.5% to 53.0% on Rationality and from 22.0% to 46.7% on Goal-Directed Social Reasoning, while Moral/Conv. improves from 52% to 64%. Similar gains occur for Qwen3.5-27B ($21.5\% \rightarrow 49.0\%$, $30.7\% \rightarrow 44.7\%$, $32\% \rightarrow 44\%$) and Qwen3.5-9B ($12.0\% \rightarrow 47.0\%$, $6.0\% \rightarrow 24.0\%$, $8\% \rightarrow 44\%$), respectively. Although reasoning does not improve every individual condition, it substantially improves pair-level reliability across all three model pairs.

A.8.3 HIGH MORAL PERFORMANCE IS NOT UNIFORM ACROSS MORAL REASONING TASKS

Moral Reasoning is the strongest domain overall, but the high aggregate performance is not shared equally by its three subdomains. Responsibility Taking is close to ceiling for several models:

Table 7: Class-conditional single-video accuracy (%) for all evaluated models. Each domain contains equally many agent-favourable and agent-unfavourable examples. For Rationality, the favourable class is *Rational*; for Social Reasoning, it is *Help*; and for Moral Reasoning, it comprises *Unintentional*, *Takes Responsibility*, and *Conventional Transgression*.

	Rationality		Social		Moral	
Model	Rational	Irrational	Help	Hinder	Favourable	Unfavourable
InternVL3.5-8B Inst	78.5	33.0	88.0	13.3	96.1	35.1
LLaVA-OV-2-8B Inst	75.0	32.0	89.3	15.3	97.4	40.3
Keye-VL-1.5-8B Think	57.0	63.5	70.7	38.7	93.5	35.1
Qwen3.5-9B Inst	81.5	25.5	98.7	6.7	83.1	63.6
Qwen3.5-9B Think	87.5	55.0	83.3	34.7	96.1	63.6
Qwen3.5-27B Inst	93.0	25.0	74.7	52.7	92.2	66.2
Qwen3.5-27B Think	94.0	51.0	82.0	55.3	98.7	72.7
Qwen3.6-27B Inst	94.5	41.5	92.7	27.3	94.8	70.1
Qwen3.6-27B Think	90.0	58.0	92.0	52.0	97.4	74.0
Qwen3-VL-32B Think	81.0	63.0	74.7	57.3	90.9	62.3
Gemini-3.1-Pro	84.0	77.5	72.0	57.3	96.1	87.0
Gemini-3.5-Flash	92.5	68.5	77.3	69.3	94.8	94.8
GPT5.6 Luna	95.0	43.0	84.7	50.0	100.0	74.0
Human	99.0	97.0	98.7	97.3	97.4	98.7

Qwen3.5-27B-Think reaches 100% Joint-Pair Accuracy, Gemini-3.1-Pro and Qwen3.6-27B-Think reach 96%, and Gemini-3.5-Flash reaches 93%.

Performance is substantially lower on Moral-vs.-Conventional judgments for many models. Gemini-3.5-Flash reaches 84% Joint-Pair Accuracy, Gemini-3.1-Pro reaches 68%, and GPT5.6 Luna reaches 48%, while Qwen3-VL-32B-Think reaches 16%, Keye-VL-1.5-8B-Think reaches 8%, and Qwen3.5-27B-Instruct reaches 32%.

Intentionality also shows considerably larger variation across models than Responsibility Taking. Thus, strong aggregate performance on the Moral domain should not be interpreted as uniform reliability across different kinds of moral judgment.

A.8.4 Open reasoning models are competitive with proprietary models on several subdomains

The strongest proprietary model is not uniformly better across SMORE. Open reasoning models match or exceed all three proprietary models on several subdomains.

The clearest case is Open/Close. Qwen3.6-27B-Think obtains 81% Paired accuracy, 81% Single accuracy, and 68% Joint-Pair Accuracy. The best proprietary scores on the same measures are 70%, 77%, and 57%, respectively. Qwen3.5-27B-Think also reaches 66% Joint-Pair Accuracy on this condition.

On True Belief, Qwen3.5-27B-Think obtains 98% Paired accuracy, 94% Single accuracy, and 88% Joint-Pair Accuracy, matching or exceeding the proprietary models. On Responsibility Taking, the same model reaches 100% in both Single accuracy and Joint-Pair Accuracy.

Find provides another notable comparison. Qwen3.6-27B-Think and Qwen3.5-27B-Think reach 88% and 83% Paired accuracy, respectively, compared with 78% for the strongest proprietary model, although their Joint-Pair Accuracy remains low.

These results do not indicate a general advantage for open models. Rather, they show that model rankings depend strongly on the reasoning domain and evaluation protocol, and that open reasoning models are already competitive with proprietary models on several comparisons.

Table 8: **Evaluation of human-behavior-specialized MLLMs on SMORE (%).** We evaluate OmniSapiens-7B-RL, OmniSapiens-7B 2.0, and HumanOmniV2, with Qwen2.5-Omni-7B included as a matched base-model reference. We report Paired accuracy (P), Single-video accuracy (S), and Joint-Pair Accuracy (J), as defined in Section 2.2. Human scores use three independent annotations per question from a pool of 65 annotators. Values shown in red indicate Joint-Pair Accuracy below the 25% chance level.

	Rationality														
	True Belief			False Belief			Believed Safe			Believed Risky			Overall		
Model	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J
Human	**100.0**	**98.0**	**96.0**	**100.0**	**95.0**	**90.0**	**100.0**	**99.0**	**98.0**	**98.0**	**100.0**	**100.0**	**99.5**	**98.0**	**96.0**
Chance	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0
Qwen2.5-Omni-7B Think	54.0	73.0	50.0	46.0	53.0	24.0	46.0	51.0	4.0	56.0	50.0	0.0	50.5	56.8	19.5
OmniSapiens-7B-RL Think	52.0	72.0	48.0	48.0	56.0	26.0	54.0	47.0	4.0	56.0	51.0	2.0	52.5	56.5	20.0
OmniSapiens-2.0 Think	52.0	61.0	22.0	48.0	45.0	4.0	60.0	43.0	2.0	46.0	52.0	14.0	51.5	50.3	10.5
HumanOmniV2	54.0	63.0	34.0	46.0	48.0	16.0	62.0	54.0	20.0	36.0	53.0	20.0	49.5	54.5	22.5

	Goal-Directed Social Reasoning												Moral Reasoning								
	Find			Open/Close			On/Off			Overall			Int.			Resp.			Moral/Conv.		
Model	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J	P	S	J
Human	**100.0**	**96.6**	**93.2**	**97.9**	**98.9**	**97.9**	**100.0**	**98.9**	**97.7**	**99.3**	**98.0**	**96.0**	**100.0**	**98.0**	**96.0**	**100.0**	**98.1**	**96.3**	**100.0**	**98.0**	**96.0**
Chance	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0
Qwen2.5-Omni-7B Think	45.8	50.0	6.8	68.1	56.4	21.3	50.0	50.0	15.9	54.0	52.0	14.0	56.0	56.0	12.0	85.2	66.7	40.7	84.0	52.0	4.0
OmniSapiens-7B-RL Think	52.5	54.2	13.6	72.3	62.8	36.2	50.0	62.5	31.8	58.0	59.3	26.0	60.0	54.0	8.0	81.5	66.7	48.1	92.0	56.0	16.0
OmniSapiens-2.0 Think	49.2	50.8	1.7	66.0	68.1	40.4	34.1	61.4	29.5	50.0	59.3	22.0	56.0	60.0	20.0	88.9	70.4	44.4	80.0	50.0	0.0
HumanOmniV2	47.5	55.9	22.0	66.0	52.1	10.6	63.6	48.9	9.1	58.0	52.7	14.7	72.0	66.0	32.0	74.1	59.3	18.5	64.0	50.0	0.0

A.8.5 Scaling Qwen models does not remove the efficiency shortcut

Comparing Qwen3.5-9B with Qwen3.5-27B, which share a model generation, scale improves Goal-Directed Social Reasoning in both configurations (Joint-Pair Accuracy $6.0\% \rightarrow 30.7\%$ for Instruct and $24.0\% \rightarrow 44.7\%$ for Think), but yields smaller gains on Rationality ($12.0\% \rightarrow 21.5\%$ and $47.0\% \rightarrow 49.0\%$). These Rationality gains come from the belief conditions; on the path conditions, scale does not help. Joint-Pair Accuracy on Believed Safe changes from 6% to 4% (Instruct) and 22% to 28% (Think), and on Believed Risky from 8% to 4% and 26% to 20%. For the Instruct models, the larger model shows a stronger efficiency shortcut: Paired accuracy rises from 56% to 96% on Believed Safe but falls from 58% to 20% on Believed Risky. Reasoning also matters more than scale on Rationality, where Qwen3.5-9B-Think (47.0%) far exceeds Qwen3.5-27B-Instruct (21.5%). Across generations, parameter count is not the dominant factor: the largest model, Qwen3-VL-32B-Think, is comparable to Qwen3.5-27B-Think on Rationality (50.0% vs. 49.0%) and lower on Goal-Directed Social Reasoning (34.0% vs. 44.7%).

A.8.6 Human-Behavior-Specialized MLLMs Show Mixed Transfer to SMORE

We additionally evaluate three MLLMs explicitly specialized for human-behavior understanding: OmniSapiens-7B-RL, introduced with the Human Behavior Atlas (Ong et al., 2026b); OmniSapiens-7B 2.0 (Ong et al., 2026a); and HumanOmniV2 (Yang et al., 2025b). Since these models build on Qwen2.5-Omni-7B (Xu et al., 2025), we also evaluate Qwen2.5-Omni-7B as a matched base-model reference.

Table 8 reports the complete results. Human-behavior specialization does not lead to consistent improvements across the three reasoning domains. On Belief-Conditioned Rationality, Qwen2.5-Omni-7B obtains 19.5% Joint-Pair Accuracy, compared with 20.0% for OmniSapiens-7B-RL, 10.5% for OmniSapiens-7B 2.0, and 22.5% for HumanOmniV2. Thus, all four models remain below the 25% Joint-Pair chance level on Rationality.

The clearest gains appear in Goal-Directed Social Reasoning. OmniSapiens-7B-RL reaches 26.0% Joint-Pair Accuracy and OmniSapiens-7B 2.0 reaches 22.0%, compared with 14.0% for their Qwen2.5-Omni-7B base model. HumanOmniV2 reaches 14.7%. The Moral subdomains show a similarly mixed pattern. For example, HumanOmniV2 improves substantially over Qwen2.5-Omni-7B on Intentionality (32.0% vs. 12.0% Joint-Pair Accuracy), but performs worse on Responsibility Taking

(18.5% vs. 40.7%). OmniSapiens-7B-RL improves Responsibility Taking from 40.7% to 48.1% and Moral/Conventional Transgressions from 4.0% to 16.0%, while its Intentionality performance decreases from 12.0% to 8.0%.

Taken together, specialization for broad human-behavior understanding can improve particular SMORE conditions, especially within Goal-Directed Social Reasoning, but these gains do not transfer uniformly across reasoning domains. In particular, the large failures on Belief-Conditioned Rationality persist even for models explicitly trained for human-behavior understanding. This suggests that the difficulties exposed by SMORE are not explained solely by evaluating general-purpose MLLMs.

A.9 Reasoning Trace Analysis

To understand the reasoning patterns associated with model errors, we complement the behavioural results with an analysis of the reasoning traces of Qwen3.6-27B and Qwen3.5-27B and the reasoning summaries of GPT-5.6 Luna and Gemini 3.1 Pro on Goal-Directed Social Reasoning and Belief-Conditioned Rationality. The Qwen models expose generated reasoning traces directly, whereas GPT-5.6 Luna and Gemini 3.1 Pro provide reasoning summaries.

Our goal is to move beyond identifying *whether* models fail and examine *how* their reasoning differs when they do. We first identify recurring reasoning patterns in the generated traces and summaries, then quantify how often these patterns occur across models and evaluation settings. Finally, we examine whether the presence of each pattern is associated with model correctness by comparing its prevalence in correct and incorrect responses. This allows us to distinguish recurring reasoning tendencies from patterns that are specifically associated with model errors.

Pattern discovery and annotation. We first identify recurring, behavior-specific reasoning patterns from a subset of the traces and define these patterns before examining their relationship with model correctness. We then use GPT-5.6 Luna as a correctness-blind annotator to determine which of the predefined patterns are present in each trace. For each example, the annotator receives only a neutral task description, the question shown to the evaluated model, the model's final answer, and its reasoning trace. Ground truth, correctness, subclass, item identifiers, and agent roles are withheld and joined only after annotation. Each identified pattern is required to be supported by a verbatim span from the trace.

We focus on four recurring reasoning patterns:

- **Affordance substitution.** The reasoning treats a locally useful-looking action—such as opening, holding, switching on, picking up, or otherwise making an object accessible—as evidence of helping without tracking whether the action ultimately advances the recipient's goal. For example, picking up or carrying the requested object may be treated as sufficient evidence of helping even when the reasoning does not follow where the object ultimately ends up.
- **Actual-world or outcome-based reasoning.** The reasoning judges rationality using the object's actual location, visible environmental conditions, or the observed success or failure of the action, rather than evaluating the action from the agent's stated belief. For example, an action may be judged rational because the object was ultimately found or because the observed route was in fact safe.
- **Speculative reinterpretation.** The reasoning explains away an apparent mismatch between the agent's stated belief and its action by introducing an unsupported alternative interpretation of the environment. This can include assuming that a room, object, label, or environmental condition has a broader or different meaning than what is explicitly established in the example.
- **Belief–action matching.** The reasoning uses agreement between the agent's stated belief and its observed action as the primary test of rationality, without fully considering route efficiency or alternative actions. For example, visiting the location where the agent believes an object to be may be treated as sufficient evidence of rationality, while an action that does not literally match the stated belief may be treated as irrational.

MODEL	$P(c \mid \text{wrong})$	$P(c \mid \text{correct})$	Δ
Qwen3.6-27B	89%	23%	+66
Qwen3.5-27B	84%	21%	+63
GPT-5.6 Luna	82%	72%	+10
Gemini 3.1 Pro	40%	31%	+9

Table 9: **Affordance substitution on single-video FIND examples.** The pattern is substantially more frequent in incorrect than correct traces for both Qwen models; both differences remain significant after FDR correction.

SETTING	MODEL	ACTUAL-WORLD OUTCOME	SPECULATIVE REINTERPRETATION	BELIEF–ACTION MATCHING
Single	Qwen3.6-27B	94 / 64	60 / 34	42 / 70
	Qwen3.5-27B	90 / 65	59 / 28	45 / 68
	GPT-5.6 Luna	54 / 51	15 / 8	25 / 48
	Gemini 3.1 Pro	50 / 40	12 / 5	21 / 58
Paired	Qwen3.6-27B	93 / 49	56 / 22	22 / 63
	Qwen3.5-27B	64 / 45	56 / 15	24 / 64
	GPT-5.6 Luna	71 / 48	36 / 6	7 / 32
	Gemini 3.1 Pro	21 / 26	12 / 6	35 / 74

Table 10: **Reasoning patterns in Belief-Conditioned Rationality.** Each entry reports $P(\text{code} \mid \text{wrong})/P(\text{code} \mid \text{correct})$ (%).

GPT-5.6 Luna codes all 1,613 Goal-Directed Social Reasoning traces and all 2,293 Belief-Conditioned Rationality traces. We compare $P(\text{code} \mid \text{wrong})$ and $P(\text{code} \mid \text{correct})$ to distinguish patterns that are merely frequent from those specifically associated with model errors.

A.9.1 GOAL-DIRECTED SOCIAL REASONING

The clearest error-associated pattern is *affordance substitution*: models can treat an immediately useful-looking action as evidence of helping without tracking whether it ultimately advances the recipient's goal. This effect is strongest in FIND. In both videos of a pair, the responding agent picks up the requested object, and only its final destination determines whether it helps or hinders: it either brings the object closer to the requester, thereby helping, or moves it farther away, thereby hindering.

For both Qwen models, incorrect traces far more often treat picking up the object as sufficient evidence of helping, without following whether it is ultimately brought closer to or moved away from the requesting agent. Affordance substitution occurs in 84–89% of incorrect traces but only 21–23% of correct traces, and both differences remain significant after correction for multiple comparisons (Table 9). GPT-5.6 Luna shows this pattern frequently in both incorrect and correct traces (82% versus 72%), so it does not distinguish its errors, while Gemini 3.1 Pro shows it less often and with only a small difference between the two (40% versus 31%).

A.9.2 BELIEF-CONDITIONED RATIONALITY

Rationality errors show a different pattern. Although the task asks whether an action is rational *given the agent's stated belief*, incorrect traces more often judge the action by the actual state of the world or by whether it succeeds, rather than from the agent's stated belief.

Actual-world or outcome-based reasoning is more frequent in incorrect traces for both Qwen models in both the single-video and paired settings (Table 10). The difference is particularly large for Qwen3.6-27B: 94% versus 64% in the single-video setting and 93% versus 49% in the paired setting. For GPT-5.6 Luna, the association appears in the paired setting only (71% versus 48%, compared with 54% versus 51% in the single-video setting). Gemini 3.1 Pro shows no consistent difference: a modest difference in single-video evaluation (50% versus 40%) and a slight reversal in the paired setting (21% versus 26%).

The subclass analysis clarifies when actual-world reasoning becomes problematic. The association is concentrated in FALSE BELIEF, where the belief and the world diverge. Across these examples, incorrect traces more often reason from the object's actual location or the observed outcome rather than from the agent's stated belief. When the belief instead agrees with reality, as in TRUE BELIEF, this relationship can weaken or even reverse. Thus, the failure is not simply the use of observable information, but relying on it when it conflicts with the perspective from which rationality should be evaluated.

The Qwen models additionally tend to explain away belief–action mismatches with unsupported reinterpretations of rooms, objects, or environmental conditions (*speculative reinterpretation*; 59–60% of incorrect versus 28–34% of correct single-video traces, and 56% versus 15–22% in the paired setting). This pattern is concentrated in FALSE BELIEF examples, where such reinterpretations can allow the reasoning to reconcile the observed action with the actual world rather than the agent's stated belief.

In contrast, one pattern holds for all four models in both settings: correct traces more often explicitly check whether the action is consistent with what the agent believes. *Belief–action matching* appears in 48–70% of correct versus 21–45% of incorrect single-video traces, and in 32–74% versus 7–35% of paired traces. Although this code was defined to capture a potentially incomplete test of rationality, one that does not fully consider route efficiency or alternative actions, explicitly grounding the judgment in the agent's stated belief is the pattern most consistently associated with correct responses.

A.10 ROBUSTNESS TO ANSWER-OPTION ORDER

In the default format, the agent-favourable answer is Option 1 in Rationality, Social Reasoning, and two of the three Moral subclasses (Option 2 in Intentionality), so the asymmetry in Section 3.3 could in principle reflect a preference for the first-listed option. To test this, we re-evaluate seven models with the two option texts exchanged while the labels (1 and 2) stay in place; prompts, frame sampling, decoding, and scoring are otherwise identical.

Table 11 reports class-conditional single-video accuracy under both orders and their counterbalanced average. After averaging, the favourable class is more accurate than the unfavourable class in every model–domain cell (19 of 19). The advantage also holds in the swapped order alone, where the favourable answer is listed in the other position, so the asymmetry is not explained by option position. Position does affect its magnitude for some models. For Keye-VL-1.5-8B Think on Rationality, the default order slightly reverses the gap (57.0 vs. 63.5), while the counterbalanced average shows a clear favourable advantage (69.8 vs. 44.8). For Qwen3.6-27B Inst on Social, the gap is much smaller in the swapped order (62.0 vs. 59.3) than in the default order (92.7 vs. 27.3).

A.11 EXTENDED REPRESENTATIONAL ANALYSIS

A.11.1 LINEAR PROBING SETUP

Linear probes provide a lightweight diagnostic for determining whether a target distinction can be recovered from a model's internal representations (Alain & Bengio, 2016; Hewitt & Liang, 2019; Gottesman & Geva, 2024). Recent work has similarly analyzed internal representations of vision-language models and has shown that information relevant to downstream behavior can vary substantially across representation type, token position, and model depth (Jiang et al., 2025; Kogilathota et al., 2026).

Features. We probe representations from three stages of video MLLM processing. The *Pooled ViT Tokens* feature, $\mathbf{h}^{(\ell)}_{\text{vis}}$, is obtained by mean-pooling the vision-encoder patch-token states and captures visual information before multimodal projection and language-model processing. The *Last Video Token* feature, $\mathbf{h}^{(\ell)}_{\text{vid}}$, is the language-model hidden state at the final position of the video-token sequence, capturing visual information after it has been processed within the language model but before the task query. Finally, the *Last Query Token* feature, $\mathbf{h}^{(\ell)}_{\text{qry}}$, is the hidden state at the final position of the task query, where both the video and query have been jointly contextualized

Table 11: **Class-conditional single-video accuracy under both option orders (%).** Fav./Unfav.: accuracy on the agent-favourable and agent-unfavourable class, as defined in Section 3.3. The favourable class is Rational for Belief-conditioned Rationality, Help for Goal-directed Social Reasoning, and Unintentional, Takes Responsibility, or Conventional Transgression for Moral Reasoning. Orig.: default order; Swap.: option texts exchanged; Avg.: average of the two.

	Fav.			Unfav.		
Model	Orig.	Swap.	Avg.	Orig.	Swap.	Avg.
Belief-conditioned Rationality						
Keye-VL-1.5-8B Think	57.0	82.5	**69.8**	63.5	26.0	**44.8**
Qwen3.5-27B Think	94.0	96.0	**95.0**	51.0	43.0	**47.0**
Qwen3.6-27B Inst	94.5	91.0	**92.8**	41.5	42.0	**41.8**
Qwen3.6-27B Think	90.0	93.5	**91.8**	58.0	52.5	**55.2**
Gemini-3.5-Flash	92.5	96.0	**94.2**	68.5	63.5	**66.0**
GPT5.6 Luna	95.0	96.5	**95.8**	43.0	40.0	**41.5**
Goal-directed Social Reasoning						
Keye-VL-1.5-8B Think	70.7	76.7	**73.7**	38.7	34.0	**36.3**
Qwen3.5-27B Think	82.0	86.0	**84.0**	55.3	48.0	**51.7**
Qwen3.6-27B Inst	92.7	62.0	**77.3**	27.3	59.3	**43.3**
Qwen3.6-27B Think	92.0	92.0	**92.0**	52.0	50.0	**51.0**
Gemini-3.1-Pro	72.0	83.3	**77.7**	57.3	70.7	**64.0**
Gemini-3.5-Flash	77.3	81.3	**79.3**	69.3	65.3	**67.3**
GPT5.6 Luna	84.7	82.7	**83.7**	50.0	48.7	**49.3**
Moral Reasoning						
Qwen3.5-27B Think	98.7	97.4	**98.1**	72.7	83.1	**77.9**
Qwen3.6-27B Inst	94.8	94.8	**94.8**	70.1	70.1	**70.1**
Qwen3.6-27B Think	97.4	97.4	**97.4**	74.0	75.3	**74.7**
Gemini-3.1-Pro	96.1	100.0	**98.1**	87.0	89.6	**88.3**
Gemini-3.5-Flash	94.8	98.7	**96.8**	94.8	83.1	**89.0**
GPT5.6 Luna	100.0	97.4	**98.7**	74.0	93.5	**83.8**

immediately before response generation. We probe only these pre-generation states and do not use generated reasoning or answer tokens.

Layer sampling. To examine how decodability evolves with depth, we extract each feature at $\ell \in \{L/4, L/2, 3L/4, L\}$. Visual features are sampled from the corresponding layers of the vision encoder, while video- and query-token features are sampled from the language model, with L denoting the number of layers in the respective component.

Probe training and evaluation. We perform this analysis on five open-source models: InternVL3.5-8B Inst, LLaVA-OV-2-8B Inst, Qwen3.5-9B Think, Qwen3.5-27B Think and Qwen3.6-27B Think. We train ℓ_2-regularized logistic regression probes on the extracted features. Each reasoning domain is probed separately, with a distinct probe trained for each feature and layer.

We use five-fold cross-validation, assigning both videos from each minimal video pair to the same fold so that one member of a pair cannot appear in training while the other is used for evaluation. Within each training split, we use three-fold cross-validation to select the regularization strength. We report pooled out-of-fold accuracy, such that every example is evaluated exactly once by a probe trained without that example. Because the evaluation sets are balanced by construction, chance performance is 50%.

A.11.2 Moral Reasoning Probes

We apply the setup of Appendix A.11.1 to the three Moral Reasoning subdomains, probing each separately (Figure 11). As in the main text, the probe target is the benchmark label itself: Moral versus Conventional for moral–conventional transgressions, Intentional versus Unintentional for intentionality, and Takes versus Does Not Take Responsibility for responsibility taking. Across all subdomains and models, the Last Query Token is the most decodable feature, reaching 92–96% on

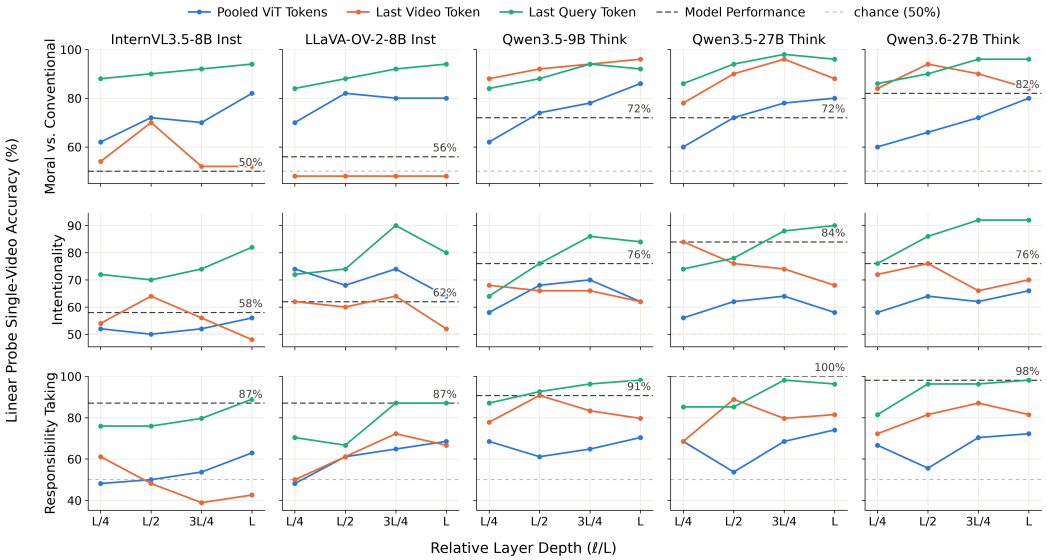

Figure 11: **Linear decodability for Moral Reasoning.** Probe accuracy for Moral vs. Conventional (top), Intentionality (middle), and Responsibility Taking (bottom). The light dashed line marks chance (50%); the dark dashed line marks the model's single-video accuracy.

Moral/Conventional, 80–92% on Intentionality, and 87–98% on Responsibility Taking at the final layer. On Moral/Conventional and Intentionality, these probes exceed single-video accuracy for all five models, most strikingly for InternVL3.5-8B and LLaVA-OV-2-8B on Moral/Conventional, whose probe accuracy reaches 94% despite behavioral accuracy of 50% and 56%. On Responsibility Taking, where behavior is already near ceiling, probes roughly match it.

The subdomains differ in where the distinction emerges. Moral/Conventional is already decodable from Pooled ViT Tokens (80–86% at the final layer). Intentionality, remains weakly decodable from visual features and emerges mainly at the Last Query Token. Last Video Token features are strong for the Qwen models but near chance for InternVL3.5-8B and LLaVA-OV-2-8B, suggesting that in these models the distinction becomes decodable only once the query is read.