SOCIALPRISM: MULTI-DIMENSIONAL GENERATIVE REWARD MODELING FOR VIDEO SOCIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video social reasoning requires models to interpret dynamic, complex, and often incongruent multimodal signals to infer hidden mental states. Existing reward modeling approaches in reinforcement learning predominantly rely on blunt scalar scores, which collapse complex social reasoning processes into coarse feedback and make it difficult for models to identify or correct specific cognitive failures. To address this bottleneck, we introduce SocialPrism, a multi-dimensional and generative reward framework that decomposes holistic social judgments into five fine-grained cognitive dimensions. We construct SocialPrism-Data to provide multi-dimensional scores and rationales for reward model training, and SocialPrism-Bench, a human-verified benchmark for reward evaluation. Furthermore, we propose Gen-SocialPrism-7B, a generative reward model trained with a two-stage curriculum learning strategy. On SocialPrism-Bench, it achieves the lowest overall score prediction error among the compared reward models. It also outperforms standard binary-reward baselines in post-training and improves test-time reward-guided selection over standard sampling baselines.

1 INTRODUCTION

Consider the example video scenario illustrated in Figure 1: judging solely from the woman's verbal response, one might conclude that she strongly dislikes the man and unequivocally rejects his romantic confession. However, her subtle facial expressions as the man leans in to kiss her reveal an underlying emotional conflict hidden beneath her harsh words, suggesting that she is, in fact, emotionally attracted to him. The ability to infer such hidden mental states from incongruent multimodal signals is a hallmark of social intelligence, commonly formalized as Theory of Mind (ToM) (Premack & Woodruff, 1978). Since real-world human interactions are inherently multimodal, dynamic, and socially nuanced, video social reasoning has emerged as a critical yet highly challenging frontier for Large Multimodal Models (LMMs).

To equip models with sophisticated social reasoning capabilities, the community has increasingly turned to post-training strategies such as reinforcement learning fine-tuning (RLFT) (Zhao et al., 2025; Yang et al., 2025; Niu et al., 2026). However, critical bottlenecks remain in the **reward modeling** stage. Firstly, although reward modeling has become a thriving research area in the language domain (Wang et al., 2024a;b), reward models tailored to complex video reasoning remain scarce. Secondly, existing reward mechanisms are often overly coarse, relying primarily on scalar rewards or vague rationales that fail to capture the nuanced cognitive processes underlying video social reasoning. When a model fails to interpret a socially complex interaction involving conflicting multimodal cues (e.g., hostile verbal expressions contradicting subtle visual signals of attraction in Figure 1), a single low reward cannot explain **why** the reasoning failed. Even when accompanied by a brief rationale, the feedback often remains superficial. By collapsing the complex chain of social reasoning into a single scalar reward or an underspecified explanation, existing reward mechanisms entangle distinct cognitive processes, making it difficult to evaluate intermediate reasoning steps, limiting their applicability to general audio-visual understanding tasks, and constraining their ability to foster deep and robust social reasoning capabilities.

To break this bottleneck, we introduce **SocialPrism**, a multi-dimensional generative reward framework as illustrated in Figure 1. SocialPrism decouples holistic social judgments into five fine-grained cognitive dimensions, ranging from foundational perception to deep mental state inference. By

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.

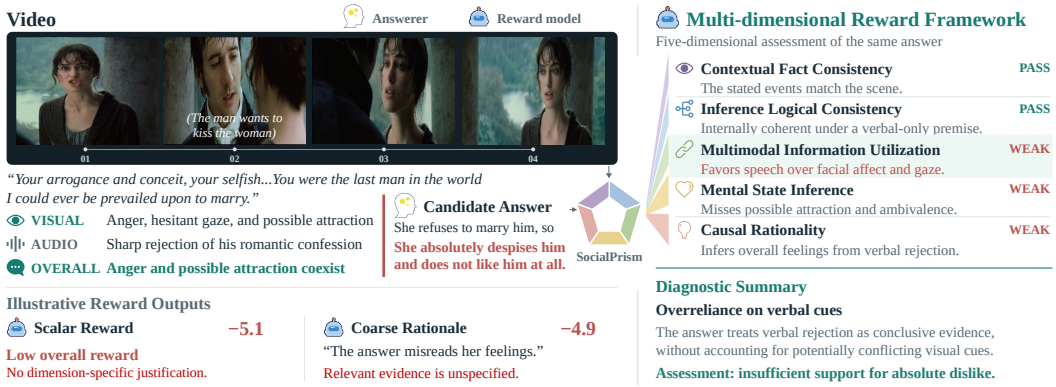


Figure 1: Overview of our multi-dimensional reward framework–SocialPrism and illustrative comparison of reward frameworks. Unlike traditional scalar rewards or vague rationales that fail to explain reasoning errors in complex social scenarios, our framework provides fine-grained, multi-dimensional rationales pinpointing the exact source of failures.

generating both discrete scores and natural-language rationales for each dimension, it provides structured feedback for inspecting the sources of reasoning failures. To support training and evaluation, we construct SocialPrism-Data, a large-scale collection of multi-dimensional scoring supervision, and SocialPrism-Bench, a benchmark with human-verified scores. Our data pipeline also produces a question-answering (QA) dataset VSR-8K and a preference dataset SocialPref-8K. We train Gen-SocialPrism-7B on SocialPrism-Data using a two-stage curriculum learning strategy. Experiments on SocialPrism-Bench show that it predicts human-verified scores more accurately than the compared reward models. It also improves **post-training** with reward feedback and **test-time inference** with reward-guided selection.

In summary, our contributions are threefold: **(1) A Novel Evaluation Framework:** We propose SocialPrism, a multi-dimensional reward evaluation framework that decomposes video social reasoning evaluation into five distinct cognitive dimensions, moving beyond the limitations of traditional reward models. **(2) Structured Training Data and a Human-verified Benchmark:** We introduce SocialPrism-Data for reward model training and SocialPrism-Bench for evaluating reward predictions, alongside the QA dataset VSR-8K and preference dataset SocialPref-8K. **(3) A Pioneering Generative Reward Model:** We develop Gen-SocialPrism-7B, which demonstrates effectiveness in providing reliable reward signals that boost performance in both post-training and test-time scenarios.

2 Related Work

2.1 Video Social Reasoning

Social intelligence is a cornerstone of human cognition. Early computational efforts to model social reasoning primarily focused on constrained environments, such as 2D virtual worlds (Wang et al., 2021; Huang et al., 2024), card games (Fuchs et al., 2021), and simulated indoor settings (Puig et al., 2023; Cao et al., 2024). Extensive research has investigated the inherent Theory of Mind (ToM) capabilities of Large Language Models (LLMs) in text-based scenarios (Shapira et al., 2023; Kim et al., 2023; He et al., 2023; Gu et al., 2024). However, real-world social interactions are inherently multimodal and dynamic. Consequently, video social reasoning has emerged as a critical research frontier. To evaluate mental state inference and social scene understanding, the community has introduced benchmarks, including MMToM-QA (Jin et al., 2024), SocialIQ (Zadeh et al., 2019), the SocialIQ 2.0 (Wilf et al., 2023), IntentBench (Yang et al., 2025), and R^3-Bench (Niu et al., 2026).

Beyond benchmarking, various training paradigms have been proposed to enhance model performance on these tasks. Early methods include CaVIR (Li et al., 2023), which leverages contrastive learning to learn intent representations, and BIP-ALM (Jin et al., 2024), which formulates a neuro-symbolic approach to infer beliefs and intents. More recently, the application of post-training strategies to LMMs has yielded improvements in video social reasoning (Yang et al., 2025; Zhao et al., 2025; Niu et al., 2026). Despite these advances, current approaches predominantly rely on rule-based

reward functions to provide training signals. Such rigid formulations fail to evaluate the intermediate reasoning steps, thereby limiting their capacity to cultivate deep social reasoning.

2.2 Reward Models

Reward models provide a flexible and scalable mechanism for evaluating generated responses. They are instrumental in both reinforcement learning (RL) pipelines (Yang et al., 2024) and test-time computation strategies (Cobbe et al., 2021; Snell et al., 2024), effectively enhancing the reasoning capabilities of foundation models. Existing reward models can be broadly categorized into two paradigms: scalar reward models, which output a single numerical score (Yang et al., 2024; Zang et al., 2025; Ji et al., 2025; Fan et al., 2025), and generative reward models, which produce both a detailed rationale and a score (Wang et al., 2025; Shao et al., 2025; Zhang et al., 2024a).

While reward modeling has seen rapid adoption in the text domain, its exploration within the video domain remains scarce. Although recent generative reward models like InternLM-XComposer2.5-Reward (Zang et al., 2025) and UNIFIEDREWARD (Wang et al., 2025) incorporate video understanding data during training, their application to complex, downstream video reasoning tasks is largely unexplored. We introduce SocialPrism-Data for learning multi-dimensional reward judgments in video social reasoning and SocialPrism-Bench for evaluating them against human-verified annotations. We also propose Gen-SocialPrism-7B, a generative reward model tailored to evaluate and enhance reasoning processes within this domain.

3 SocialPrism: A Multi-Dimensional Generative Framework

SocialPrism is a multi-dimensional reward framework that decouples a model's response into fine-grained, per-dimension scores and explanatory rationales. Its design is motivated by the observation that video social reasoning involves a hierarchy of capabilities, from low-level perception to high-level social inference, and that failures at different levels may require different feedback.

Drawing on Theory of Mind (Premack & Woodruff, 1978), Belief-Desire-Intention (BDI) model (Bratman, 1987), and social cognition studies (Tomasello, 2010; Reisenzein, 2006; Puica & Florea, 2013; Schlaffke et al., 2015; Fan et al., 2022), we derive five complementary dimensions organized along the following capability hierarchy:

(1) Perception. Recognition of objective video facts, such as actions, scenes and dialogue, forms the basis of higher-order social reasoning. This motivates **Contextual Fact Consistency (CFC)**.

(2) Multimodal Fusion. Social meaning is conveyed through both verbal cues and non-verbal cues, such as tone and gestures. Capturing the full spectrum of social signals requires deep integration across modalities. This motivates **Multimodal Information Utilization (MIU)**.

(3) Causal and Social Inference. Beyond perception and grounding, social reasoning requires plausible explanations of observed behaviors and hidden mental states, including beliefs, intentions, and emotions. This motivates **Causal Rationality (CR)** and **Mental State Inference (MSI)**.

(4) Logical Coherence. Across all levels, the reasoning chain should remain internally consistent from factual premises to social conclusions. This motivates **Inference Logical Consistency (ILC)**.

The evaluation perspective for each dimension are presented in Table 1. Each dimension is scored on a discrete 0–3 scale with explicit criteria defining every level. SocialPrism provides a **comprehensive**

Cap. Level	Dim.	Evaluation Aspect
Perception	CFC	Faithfulness of described facts to video content
Multimodal Fusion	MIU	Integration of visual and auditory modalities
Causal & Social Inference	CR	Plausibility of causal explanations for events and mental states.
	MSI	Accuracy of observable and hidden mental state inference
Logical Coherence	ILC	Internal logical consistency of reasoning process

Table 1: The five evaluation dimensions of SocialPrism, organized by the capability hierarchy considered in this work. The rubric is provided in Section B.2.

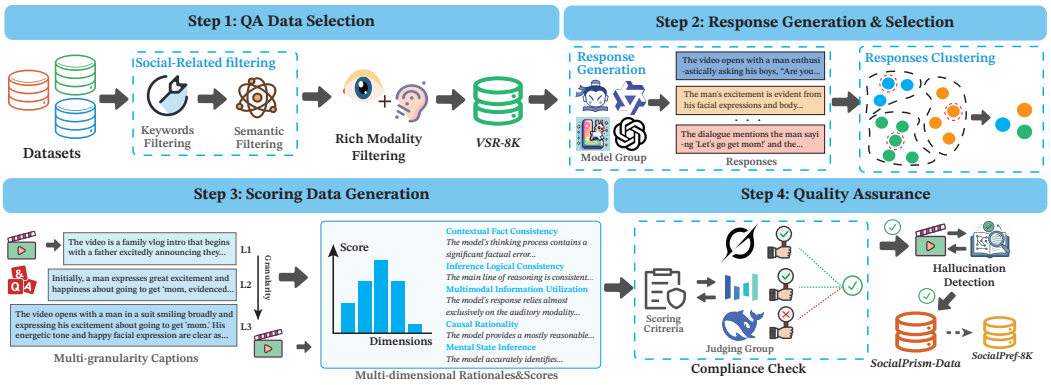


Figure 2: Overview of the dataset construction pipeline. The systematic process consists of four main steps: (1) **QA Data Selection** via social-related and modality filtering to curate relevant tasks; (2) **Response Generation & Selection** using a diverse group of multimodal models; (3) **Scoring Data Generation** across multiple dimensions, guided by multi-granularity captions; and (4) **Quality Assurance** involving rigorous compliance checks and hallucination detection to yield the final SocialPrism-Data and two derivatives (VSR-8K and SocialPref-8K).

evaluation protocol by assessing model responses along the full capability hierarchy from perception to social inference, while producing dimension-specific rationales that reveal where and why reasoning fails. Moreover, its per-dimension scores can be **flexibly** aggregated into a global reward, enabling targeted optimization of different reasoning capabilities.

Let $\mathcal{C} = \{\texttt{CFC}, \texttt{MIU}, \texttt{CR}, \texttt{MSI}, \texttt{ILC}\}$ denote the five evaluation dimensions of SocialPrism. For each dimension $c \in \mathcal{C}$, SocialPrism assigns a discrete score $s^c \in \{0, 1, 2, 3\}$ and a natural-language rationale e^c. The overall reward is obtained by a customizable aggregation function $\texttt{Agg}$:

$$r_{\texttt{sp}} = \texttt{Agg}(\{s^c\}_{c \in \mathcal{C}}). \tag{1}$$

4 Datasets

As illustrated in Figure 2, our data pipeline supports both reward model training and evaluation. It first constructs VSR-8K, generates diverse responses, and annotates them with multi-dimensional scores and rationales. Automated quality checks produce a scoring corpus from which we obtain SocialPrism-Data for training, SocialPrism-Bench through human verification of held-out samples, and SocialPref-8K through preference-pair construction. The pipeline follows three design goals: **diversity, scalability, and quality control**.

4.1 VSR-8K

We first construct a video social reasoning dataset named VSR-8K by integrating two distinct data sources: SocialIQ-2.0 Challenge (Wilf et al., 2023) and LLaVA-Video-178K (Zhang et al., 2024b). SocialIQ-2.0 features videos depicting different social scenarios, while LLaVA-Video-178K aggregates multiple datasets including NextQA (Xiao et al., 2021), Perception Test (Patraucean et al., 2023), etc. alongside a mix of long and short clips. The combination enhances the diversity of videos.

To extract relevant social reasoning content from LLaVA-Video-178K, we employ a two-stage filtering method: (i) **Keyword Filtering**: We identify candidates using keywords related to four mental states: belief, desire, intent, and emotion; and (ii) **Semantic Filtering**: We use GPT-5 nano (OpenAI, 2025a) to select semantically relevant samples.

To ensure rich multimodal signals, we exclude videos with fewer than 10 lines of dialogue detected by Whisper (Radford et al., 2023). The final VSR-8K comprises 8,345 QA pairs across 2,143 videos.

4.2 SocialPrism-Data

We design an automated pipeline to construct scalable scoring data with diverse response sources and multiple quality-control stages. We employ the model group consisting of 11 distinct models from o4-

mini (OpenAI, 2025b), Qwen (Wang et al., 2024c; Bai et al., 2025; Xu et al., 2025), InternVL3 (Zhu et al., 2025), and LLaVA-OneVision (Li et al., 2024) series (detailed in Section B.1). It encompasses both omni-modal and vision-language models (VLMs), helping to capture a wide spectrum of response styles. For VLMs, we additionally provide audio transcripts to ensure information completeness.

Empirical evaluations reveal that the reasoning chains in the model's correct responses are frequently suboptimal. To achieve a balanced distribution of scores and correctness, we sample three correct and two incorrect responses per question, provided sufficient candidates are available. We always include o4-mini as a reliable high-quality anchor. To ensure diversity and prevent shortcut learning, the remaining candidates are selected via independent clustering on the correct and incorrect subsets.

To construct scalable scoring supervision, we implement an automated evaluation pipeline with multiple quality-control stages using strong LMMs. We employ Gemini 2.5 Pro (Comanici et al., 2025), which excels in video understanding and instruction following, as the primary evaluator. We decouple the scoring task into "video understanding" and "response scoring" to reduce complexity, prevent subtask interference, and improve evaluation reliability. In the first stage, Gemini 2.5 Pro extracts relevant, multi-granular video captions based on the corresponding questions and correct options. In the second stage, it uses these captions alongside the original video as joint context to dynamically filter essential information and accurately score responses.

Finally, we introduce a multi-stage mechanism to encourage rubric adherence and filter likely video hallucinations. First, we employ Seed1.8 (Seed, 2026), Grok 4.1 Fast, and DeepSeek-Reasoner (Liu et al., 2025)—models distinguished by their exceptional logical reasoning and discriminative capabilities—as judges to assess whether the rationales and scores across all dimensions conform to the predefined rubrics, retaining only samples that win a majority vote. Next, Gemini 3 Flash verifies consistency between the scoring data and the video to filter out hallucinated samples. This screening yields a corpus of 37,087 automatically annotated samples, collectively form SocialPrism-Data.

4.3 SocialPref-8K

To endow the reward model with preference discrimination capabilities, we structurally transform the point-wise scoring data into pair-wise scoring data. From the automatically curated scoring corpus, we construct a preference dataset SocialPref-8K comprising 8,251 samples. Specifically, the construction process operates as follows: for any given question with at least two scoring data where the total scores across all dimensions differ, the highest-scoring response is designated as preferred, and the lowest-scoring response is designated as rejected.

4.4 SocialPrism-Bench

To evaluate reward prediction capability, we construct SocialPrism-Bench based on R^3-Bench-Hard. We use Qwen2.5-Omni-7B, InternVL3-8B and Qwen2.5-VL-72B to generate responses. Gemini 2.5 Pro then produces preliminary scores and rationales following the same two-stage annotation procedure, which are subsequently verified and corrected by human annotators. Each sample is independently reviewed by two human annotators, who examine the scores and rationales across all five dimensions. A sample is retained when both annotators judge all dimension-level scores and rationales to be correct, or when both identify the same error and agree on the same correction. This process yields SocialPrism-Bench, comprising 339 samples with human-verified scores. The study finds an agreement of 92.64% between the original automatic annotations and human judgments, with an inter-annotator agreement of 89.65%. We use SocialPrism-Bench to evaluate reward prediction against these human-verified annotations. Notably, since R^3-Bench-Hard is a more challenging benchmark, SocialPrism-Bench is also substantially more difficult.

5 Gen-SocialPrism-7B

We train our generative reward model, Gen-SocialPrism-7B, on SocialPrism-Data through a two-stage curriculum learning strategy built upon supervised fine-tuning (SFT). The core insight is to decompose the complex multi-dimensional scoring task into progressively harder sub-tasks: the model first learns to assess individual quality dimensions in isolation, then graduates to holistic multi-dimensional scoring and comparative judgment.

5.1 Problem Formulation

Given a video v, a question q, and a candidate response o, Gen-SocialPrism-7B is trained to produce structured evaluations across five dimensions.

Gen-SocialPrism-7B is parameterized as an autoregressive multimodal model π_θ that generates JSON outputs. Both stages are optimized with the standard next-token prediction objective:

$$\mathcal{L}_{\texttt{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{|y|}\log \pi_\theta(y_t \mid x, y_{<t})\right] \tag{2}$$

where x denotes the input context including system instruction, video information, question, and response and y denotes the target structured output.

5.2 Stage 1: Single-Dimension Evaluation

In the first stage, we train the model on single-dimension scoring, where each training sample requires evaluating exactly one dimension for a given response. Formally, for each response o_i and each dimension c, we construct a training instance:

$$x_i^c = [\texttt{inst}_c; v; q; o_i], \quad y_i^c = \texttt{json}(c, e_i^c, s_i^c) \tag{3}$$

where $\texttt{inst}_c$ is a dimension-specific evaluation instruction that provides the rubric for dimension c. Each response is expanded into 5 independent training instances.

This decomposition serves two purposes. First, it reduces the output complexity per sample, allowing the model to focus on mastering the evaluation criteria of each individual dimension. Second, it provides balanced supervision across all five dimensions, preventing the model from developing biases toward dimensions that are easier to assess.

5.3 Stage 2: Holistic Evaluation with Pairwise Comparison

In the second stage, we present the model with two complementary task formats using a mixed training set $\mathcal{D}_2 = \mathcal{D}_{\texttt{point}} \cup \mathcal{D}_{\texttt{pair}}$.

Point-wise multi-dimensional scoring. The model is required to evaluate all five dimensions simultaneously for a single response:

$$x_i = [\texttt{inst}_{\mathcal{C}}; v; q; o_i], \quad y_i = \texttt{json}(\{(c, e_i^c, s_i^c)\}_{c\in\mathcal{C}}) \tag{4}$$

This forces the model to jointly reason across dimensions, capturing inter-dimensional correlations (e.g., factual errors typically impair causal reasoning).

Pair-wise comparative judgment. In addition, we introduce pairwise comparison samples where the model receives two responses o_i^A and o_i^B to the same question and must produce a comparative analysis with a winner verdict:

$$x_i = [\texttt{inst}_{\texttt{pair}}; v; q; o_i^A; o_i^B], \quad y_i = \texttt{json}(\texttt{rationale}, z_i) \tag{5}$$

where $z_i \in \{\texttt{Model A}, \texttt{Model B}\}$ denotes the preferred response. The rationale is expected to explicitly compare the two responses across all dimensions before reaching a verdict.

The combined stage-2 objective is:

$$\mathcal{L}_{\texttt{Stage2}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\texttt{point}}}\left[\sum_t \log \pi_\theta(y_t \mid x, y_{<t})\right] - \mathbb{E}_{(x,y)\sim\mathcal{D}_{\texttt{pair}}}\left[\sum_t \log \pi_\theta(y_t \mid x, y_{<t})\right] \tag{6}$$

The inclusion of pairwise data enhances the model's discriminative capability: while point-wise scoring teaches absolute quality assessment, pairwise comparison teaches relative quality ranking, which is directly aligned with the downstream use of the reward model for preference optimization.

6 EXPERIMENTS

We first evaluate reward prediction on SocialPrism-Bench, including comparisons with reward model baselines and ablations and then evaluate its benefits in post-training and test-time selection.

6.1 EVALUATION ON SOCIALPRISM-BENCH

6.1.1 SETUP

Gen-SocialPrism-7B is initialized from Qwen2.5-Omni-7B Xu et al. (2025). Stage 1 trains for 1 epoch on the decomposed single-dimension data. Stage 2 trains for 3 epochs on the mixed point-wise and pair-wise data. We normalize the outputs of existing reward models to the SocialPrism scoring scale before computing the distance metrics. For InternLM-XComposer-2.5-Reward, we apply min–max normalization using bounds $[-8, 8]$. To mitigate metric bias arising from potential distribution shifts induced by different scoring criteria across reward models, we additionally report Spearman's rank correlation coefficient (ρ_s) between the predicted rewards and reference total scores to assess ranking agreement, as it is invariant to the normalization procedure. Please note that ρ_s is used only to compare models developed under SocialPrism with other baselines. For comparisons among models developed under SocialPrism, prediction-error metrics should be used instead. We also report two ablation variants: (1) "w/o Stage 1", which removes Stage 1 training to evaluate the effect of curriculum learning; and (2) "Total-score-only", which predicts only the overall score to assess the benefit of multi-dimensional modeling.

6.1.2 RESULTS AND ANALYSIS

	$\texttt{NMAE}_\texttt{d} \downarrow$	$\texttt{NMAE}_\texttt{all} \downarrow$	$\rho_s \uparrow$
UNIFIEDREWARD	–	0.3404	0.2861
InternLM-XComposer-2.5-Reward	–	0.3030	0.2276
Gen-SocialPrism-7B	**0.2354**	**0.2016**	**0.4247**
- w/o Stage 1	0.2403	0.2069	–
- Total-score-only	–	0.2071	–

Table 2: Reward prediction on the SocialPrism-Bench. $\texttt{NMAE}_\texttt{d}$ and $\texttt{NMAE}_\texttt{all}$ measure normalized prediction errors for dimension-level and total scores, respectively. ρ_s is the Spearman's rank correlation coefficient.

As shown in Table 2, Gen-SocialPrism-7B achieves the best overall reward prediction performance on SocialPrism-Bench. Compared with existing reward models, it substantially reduces $\texttt{NMAE}_\texttt{all}$ to 0.2016 and achieves a higher Spearman correlation of 0.4247, compared with 0.2861 for UNIFIEDREWARD and 0.2276 for InternLM-XComposer-2.5-Reward. These results indicate that Gen-SocialPrism-7B better aligns its predicted rewards with the human-validated reference scores in both prediction accuracy and ranking consistency.

The ablation results further validate our design choices. Removing Stage 1 increases $\texttt{NMAE}_\texttt{d}$ from 0.2354 to 0.2403 and $\texttt{NMAE}_\texttt{all}$ from 0.2016 to 0.2069, demonstrating the benefit of the proposed curriculum learning strategy. Moreover, the "Total-score-only" variant yields a higher total-score prediction error (0.2071 vs. 0.2016), suggesting that explicitly modeling multiple dimensions also benefits overall reward prediction.

6.2 GEN-SOCIALPRISM-7B FOR POST-TRAINING

6.2.1 SETUP

We evaluate on five benchmarks spanning diverse social reasoning and omni-modal understanding capabilities. All benchmarks are formulated as multiple-choice questions. **R^3-Bench-Hard** (Niu et al., 2026) evaluates challenging social reasoning. **EmoBench-M** (Hu et al., 2025) measures multimodal emotional intelligence, from which we use three video-related tasks: SD (Castro et al.,

	R^3-Bench-Hard	EmoBench-M			WorldSense	OmniVideoBench		FutureOmni
		SD	PEA	MPDER		(0,1] min	(1,5] min	
Open-Source MLLMs								
VITA-1.5-7B	–	–	–	–	36.9	31.33	27.41	–
VideoLLaMA2-7B	–	55.8	45.2	30.7	25.4	32.00	28.20	40.75
Qwen3-Omni-30B	–	–	–	–	54.0	45.78	37.03	53.05
Proprietary MLLMs								
GPT-4o	48.73	–	–	–	42.60	–	–	49.70
Gemini 1.5 Pro	39.24	–	–	–	48.00	–	–	–
Gemini 2.5 Flash	–	–	–	–	52.3	55.42	55.10	55.61
Gemini 2.5 Pro	59.18	–	–	–	65.1	57.83	64.43	57.93
Qwen2.5-Omni-7B								
-	31.01	62.8	62.8	43.0	45.40	39.16	27.99	47.48
+ CS	23.10	62.8	67.4	49.4	44.80	44.44	32.65	47.20
+ CS + RPO	25.95	55.2	71.2	46.2	46.6	46.20	32.06	49.52
+ CS + GRPO	32.91	64.0	72.2	49.4	46.97	46.20	31.18	49.90
+ CS + Gen-Social Prism-7B (Ours)	**33.23**	**67.8**	**75.2**	**51.4**	**47.79**	**49.12**	**35.59**	**52.42**

Table 3: Accuracy results (%) on social reasoning and omni-modal understanding benchmarks. "CS" denotes cold-start training. Integrating Gen-SocialPrism-7B with Qwen2.5-Omni-7B improves over the base model and is competitive with or better than standard RPO and GRPO on most tasks. Best results among post-training configurations are in **bold**.

2019), PEA (Liu et al., 2022) and MPDER (Poria et al., 2019). **WorldSense** (Hong et al., 2025), **OmniVideoBench** (Li et al., 2026) and **FutureOmni** (Chen et al., 2026) evaluate real-world omni-modal understanding. To eliminate interference from the gap between ultra-long and training videos, we select a subset of OmniVideoBench of 5 minutes or less for evaluation.

Using Qwen2.5-Omni-7B as the base policy, we investigate four post-training configurations: **(1)"+ CS (Cold-Start)".** Given the relatively limited reasoning capability of Qwen2.5-Omni-7B, we adopt cold-start training to bootstrap its reasoning capability. For each question, we use it to generate two candidate responses and employ Gen-SocialPrism-7B to select the preferred one for SFT.

"+ CS + RPO". We apply RPO (Pang et al., 2024) to the cold-started model on SocialPref-8K.

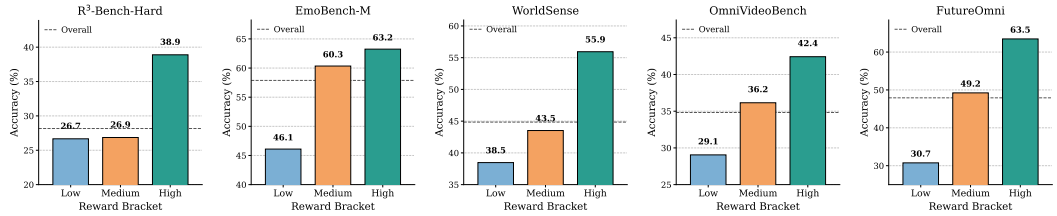
"+ CS + GRPO". We apply GRPO (Shao et al., 2024) to the cold-started model on VSR-8K with a composite reward consisting of a binary accuracy reward $r_{\texttt{acc}}$ and a format reward $r_{\texttt{fmt}}$, total reward is $r = r_{\texttt{acc}} + 0.5 \cdot r_{\texttt{fmt}}$.

"+ CS + Gen-SocPrism-7B". We apply GRPO to the cold-started model on VSR-8K where Gen-SocialPrism-7B provides a fine-grained reward signal. The reward function is detailed in Section D.1.

6.2.2 Results and Analysis

Overall post-training performance. As shown in Table 3, following cold-start training, RPO on SocialPref-8K and GRPO on VSR-8K both improve over the base model on most reported metrics, supporting the effectiveness of these datasets for post-training. Incorporating Gen-SocialPrism-7B further achieves the highest accuracy across all eight reported metrics on five benchmarks among the evaluated Qwen2.5-Omni-7B post-training configurations. Compared with the base model, the gains are particularly pronounced on EmoBench-M "PEA" and the OmniVideoBench "$(0, 1]$ min" subset, reaching 12.40% and 9.96% points, respectively. These results demonstrate the effectiveness of our post-training pipeline across social reasoning and omni-modal understanding tasks.

Contribution of the Gen-SocialPrism-7B. To examine the contribution of the reward formulation, we compare our method with "+ CS + GRPO", which shares the same cold-start initialization, GRPO algorithm, and QA training dataset. Our method outperforms it across all eight reported metrics, with notable gains of 3.80 percentage points on EmoBench-M "SD" and 4.41 points on the OmniVideoBench "$(1, 5]$ min" subset. Results demonstrate the benefit of incorporating fine-grained reward signals from Gen-SocialPrism-7B, enriching the supervision beyond binary accuracy.



(a) R^3-Bench-Hard (b) EmoBench-M (c) WorldSense (d) OmniVideoBench (e) FutureOmni

Figure 3: Accuracy of Qwen2.5-Omni-7B across different reward brackets assigned by Gen-SocialPrism-7B. Total rewards are categorized into brackets: Low, Medium, and High, corresponding to intervals $[0, |\mathcal{C}_b|]$, $(\mathcal{C}_b|, 2|\mathcal{C}_b|]$, and $(2|\mathcal{C}_b|, 3|\mathcal{C}_b|]$, respectively,where $|\mathcal{C}_b|$ denotes the number of dimensions utilized for benchmark b.

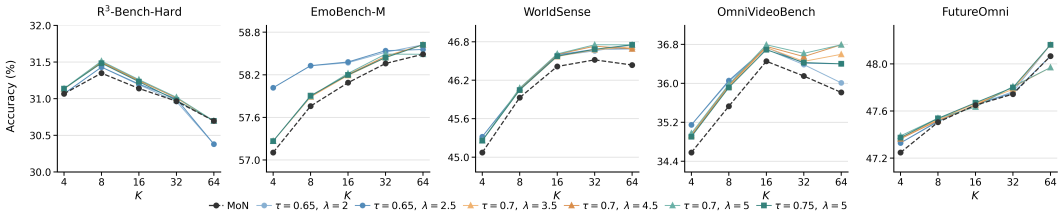


Figure 4: Accuracy comparison of MoN and Gen-SocialPrism-7B voting under different sampling budgets. For $K \in \{4, 8, 16, 32, 64\}$, accuracy is averaged over 100 random trials with fixed seed 42.

6.3 Gen-SocialPrism-7B at Test-Time

6.3.1 Setup

Using the benchmarks in Section 6.2.1, we evaluate Gen-SocialPrism-7B in two test-time scenarios: **inference** and **sampling**. In the inference scenario, we analyze how Qwen2.5-Omni-7B accuracy varies across reward brackets assigned by Gen-SocialPrism-7B. In the sampling scenario, we test whether reward-guided voting improves Qwen2.5-Omni-7B under different sampling budgets. We set the decoding temperature to $0.$ for inference and $1.$ for sampling. For omni-modal benchmarks, which typically do not involve mental states, we exclude the MSI dimension and aggregate remaining dimensions. During sampling, each candidate response casts a reward-guided vote for its parsed answer, with the detailed formulations provided in Section D.3. We evaluate sampling budgets K under multiple (τ, λ) configurations.

6.3.2 Results and Analysis

For the inference scenario, Figure 3 illustrates accuracies of Qwen2.5-Omni-7B across different reward brackets. The results clearly demonstrate a strong positive correlation between accuracy and the assigned reward brackets: in almost all benchmarks, higher reward brackets correspond to higher accuracy, and the performance in the "High" bracket consistently surpasses the "Overall" baseline, indicating that Gen-SocialPrism-7B provides an effective discriminative signal during inference.

As illustrated in Figure 4, Gen-SocialPrism-7B consistently outperforms or performs competitively with the standard majority-of-N (MoN) voting baseline under identical sampling budgets (K), yielding particularly pronounced improvements on EmoBench-M, WorldSense, and OmniVideoBench. The consistent performance gains across a diverse set of (τ, λ) configurations indicate the robustness of the reward signal from Gen-SocialPrism-7B, demonstrating that it captures reliability information well beyond simple frequency. We observe a coupling between the threshold τ and amplification weight λ: a higher τ selects fewer responses, requiring a larger λ to dominate the prediction. Crucially, achieving superior accuracy by upweighting this sparse subset proves our high-reward signals strongly align with high-quality responses.

AI USE STATEMENT

We used generative AI tools to assist in the construction of SocialPrism-Data and SocialPrism-Bench. We also used these tools to polish the manuscript, improving its grammar, clarity, and readability. The authors take full responsibility for the final content of this work, including the datasets, benchmark, and text produced with AI assistance.

ETHICS STATEMENT

This work aims to advance video social reasoning through SocialPrism-Data and SocialPrism-Bench, constructed from existing datasets. Our resources are intended for research, and we discourage their use for invasive surveillance, discriminatory profiling, or unsupported judgments about individuals. Any reuse or redistribution should respect the original data licenses, privacy, and the rights of individuals depicted in the videos. We document our data construction and evaluation procedures and disclose AI assistance to support transparency and reproducibility.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Zhihao Cao, Zidong Wang, Siwen Xie, Anji Liu, and Lifeng Fan. Smart help: Strategic opponent modeling for proactive and adaptive robot assistance in households. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18091–18101, 2024.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4619–4629, 2019.

Qian Chen, Jinlan Fu, Changsong Li, See-Kiong Ng, and Xipeng Qiu. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. *arXiv preprint arXiv:2601.13836*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *arXiv preprint arXiv:2505.17018*, 2025.

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

Andrew Fuchs, Michael Walton, Theresa Chadwick, and Doug Lange. Theory of mind for deep reinforcement learning in hanabi. *arXiv preprint arXiv:2101.09328*, 2021.

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

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

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Yizhe Huang, Anji Liu, Fanqi Kong, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Efficient adaptation in mixed-motive environments via hierarchical opponent modeling and planning. *arXiv preprint arXiv:2406.08002*, 2024.

Jiaming Ji, Xinyu Chen, Rui Pan, Conghui Zhang, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, et al. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. *arXiv preprint arXiv:2503.17682*, 2025.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

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

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Heying, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, ke wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi LI, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnivideobench: Towards audio-visual understanding evaluation for omni MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ItRYEe8E61.

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

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction*, pp. 247–258, 2022.

Lixing Niu, Jiapeng Li, Xingping Yu, Xinyi Dong, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Read the room: Video social reasoning with mental-physical causal chains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TJilJnZjpw.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025a.

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 2025b.

Richard Y Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

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

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

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. *arXiv preprint arXiv:2301.05223*, 2023.

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

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

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

Bytedance Seed. Seed1. 8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

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

Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*, 2025.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

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

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024c.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. *arXiv preprint arXiv:2111.09189*, 2021.

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

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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*, 2025.

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

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2024. *URL https://arxiv.org/abs/2408.15240*, 1, 2024a.

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

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Conclusion

In this work, we introduce SocialPrism, a multi-dimensional and generative reward framework that evaluates video social reasoning across five cognitive dimensions, providing dimension-specific scores and natural-language rationales. We construct SocialPrism-Data for reward model training and SocialPrism-Bench for evaluation, alongside VSR-8K and SocialPref-8K for QA and preference training. Building on these resources, we develop Gen-SocialPrism-7B through a two-stage curriculum for point-wise scoring and pair-wise comparison. Experiments on SocialPrism-Bench show that Gen-SocialPrism-7B consistently outperforms all compared reward models, demonstrating its effectiveness in providing reliable reward signals for video social reasoning. Further experiments show that its fine-grained rewards improve post-training beyond binary accuracy rewards, while also benefiting test-time prediction through reward-guided voting. Together, these contributions provide a framework and resources for evaluating and improving video social reasoning.

B Details of Datasets Construction

B.1 Models Used for Generating Responses

Table 4 lists the names of all models in the model group used to generate responses. We carefully select these models based on their capabilities and supported input modalities to produce a score distribution that is as balanced as possible.

Supported Modalities	Model Name
V+L	InternVL3-2B
V+L	InternVL3-8B
V+A+L	Qwen2.5-Omni-3B
V+A+L	Qwen2.5-Omni-7B
V+L	Qwen2.5-VL-3B
V+L	Qwen2.5-VL-7B
V+L	Qwen2.5-VL-72B
V+L	Qwen2-VL-2B-Instruct
V+L	Qwen2-VL-7B-Instruct
V+L	llava-onevision-qwen2-7b-ov-hf
V+L	o4-mini

Table 4: Models used for generating responses.

B.2 Prompts for Generating Scoring Data

Here are the system prompt and user prompt used for generating scoring data using Gemini 2.5 Pro. In the system prompt, we provide the model with detailed scoring criteria for each score within each dimension, which helps the model generate reasonable scoring data.

System Prompt for Scoring Data Generation

Task
You will act as an impartial judge. Your task is to meticulously evaluate the quality of a model's responses to a specific question about a video consisting of rich social interactions. The response consists of two parts: a "thinking process" and a "final answer". Here are some guidelines for evaluation:
- You are provided with predefined evaluation dimensions. For each dimension, you need to thoroughly evaluate the quality of the model's response within that dimension, and then provide reasons for your scoring and the corresponding score (0 to 3).
- All your evaluation must be grounded in the provided information about the video.
- Do not let the length or writing style of the response influence your evaluation.

- The correctness of the final answer is determined through other methods and should not be part of your evaluation. You only need to evaluate the 5 predefined dimensions.
- Your evaluation of the "Inference Logical Consistency" dimension must be based exclusively on the model's response. For other dimensions, you must evaluate the model's response in conjunction with the video context.

Predefined Evaluation Dimensions
Contextual Fact Consistency: Evaluate whether the objective information described by the model is entirely faithful to the video content.
- 0 points (Contains Serious or Numerous Errors): Contains numerous factual errors, or a critical error on a key fact, causing the analysis to be based on a flawed premise.
- 1 point (Contains Noticeable Errors): Includes one or more noticeable factual errors that could potentially affect the understanding of the events.
- 2 points (Mostly Consistent): Contains very few minor factual errors that do not affect the core understanding (e.g., slight deviation in color or count).
- 3 points (Perfectly Consistent): All descriptions of objective facts from the video (characters, actions, objects, environment, dialogue) are 100% accurate with no errors.

Inference Logical Consistency: Evaluate whether the model's response is free from internal logical contradictions. This dimension judges the structure of the argument, not its plausibility or factual accuracy. You need to evaluate the internal logical consistency of the thinking process as well as the logical consistency between the thinking process and the answer.
- 0 points (Logically Chaotic): The response is full of self-contradictions, failing to form a structurally valid argument.
- 1 point (Contains Obvious Flaws): The response contains one or more noticeable internal contradictions or logical discontinuities that undermine its structural integrity.
- 2 points (Mostly Coherent): The main line of the response is internally consistent, but there may be minor logical flaws or contradictions in secondary details.
- 3 points (Logically Tight): The entire response is logically rigorous and internally consistent from premise to conclusion, with no self-contradictions.

Multimodal Information Utilization: Evaluate whether the model effectively integrates information from both visual and auditory (dialogue) modalities.
- 0 points (Single-Modality Focus): The response relies almost exclusively on information from a single modality (typically dialogue), with minimal or superficial reference to the other.
- 1 point (Flawed Integration): The model attempts to use both modalities, but makes errors in integrating them, leading to illogical or contradictory conclusions. For example, the model attributes clues (words spoken, actions, expressions, etc.) to the wrong person.
- 2 points (Superficial Integration): The model correctly combines information from both modalities, but the integration lacks depth and misses the nuance or deeper implications of how the modalities interact.
- 3 points (Deep Integration & Insight): The response demonstrates a comprehensive understanding by weaving together visual details (e.g., character expressions, actions, environment) and dialogue to derive insights that would be impossible to obtain from either modality alone.

Causal Rationality: Evaluate the plausibility and depth of the model's explanations for why events, behaviors or mental states occur. It judges the soundness of the connective "because" link in the model's reasoning.
- 0 points (Completely Unreasonable):The explanation of causality is illogical, contradicts common sense, or is purely speculative with no basis in the evidence.
- 1 point (Partially Reasonable): The explanation contains logical leaps or is far-fetched. It overlooks more obvious causal links in favor of a less plausible one.
- 2 points (Mostly Reasonable): The main causal relationship is correctly identified and the logic is generally sound, but the explanation may be superficial, lacking depth or failing to explore more complex possibilities.
- 3 points (Insightful & Reasonable): The explanation of causality is comprehensive and

logically rigorous, offering profound insights into the causal relationships among the background, events, subtle social cues, and mental states.

Mental State Inference: Evaluate the ability to accurately and insightfully infer mental states. It evaluates the capacity to identify observable and straightforward mental states (e.g. the emotions shown on the face, the expressed beliefs), as well as the deeper, hidden motivations, temperaments, desires, beliefs, and intentions.
- 0 points (Incorrect or No Attempt): The model fails to identify any relevant mental states, makes grossly incorrect interpretations, or simply restates actions without psychological inference.
- 1 point (Poor or Minimal Inference): The model identifies some observable mental states (e.g., "sad") but with weak justification, or makes an implausible guess about hidden mental states. The inference is vague and lacks explanatory power.
- 2 points (Good or Sufficient Inference): The model accurately identifies the primary observable mental states and makes a plausible inference about hidden states, supported by direct evidence. However, the analysis may lack specificity or depth (e.g., identifies "sad" but misses the nuance of "disappointed" or "resigned").
- 3 points (Excellent or Comprehensive Inference): The model insightfully identifies specific and nuanced mental states (both observable and hidden). It provides convincing justification by integrating multiple clues and explains the complexity and implications of these internal states.

Inputs
You will receive a video, its summary, the question, the ground truth answer and the model's response.

Output Format
{
    “contextual_fact_consistency": {
        “rationale": Detailed explanation for scoring on “contextual fact consistency" dimension
        “score": The score for “contextual fact consistency" dimension
    },
    “inference_logical_consistency": {
        “rationale": Detailed explanation for scoring on “inference logical consistency" dimension
        “score": The score for “inference logical consistency" dimension
    },
    “multimodal_information_utilization": {
        “rationale": Detailed explanation for scoring on “multimodal information utilization" dimension
        “score": The score for “multimodal information utilization" dimension
    },
    “causal_rationality": {
        “rationale": Detailed explanation for scoring on “causal rationality" dimension
        “score": The score for “causal rationality" dimension
    },
    “mental_state_inference": {
        “rationale": Detailed explanation for scoring on “mental state inference" dimension
        “score": The score for “mental state inference" dimension
    },
}

User Prompt for Scoring Data Generation

Video Summary
{Summary}

Question
{Question}
Correct Answer
{Answer}

A Model's Response
{Response}

B.3 Prompts for Judging

These are the system and user prompts used to judge whether the generated scoring data meets the scoring criteria.

System Prompt for Judging

You are an expert Quality Assurance Auditor for content evaluation. You are tasked with verifying the accuracy of scoring data provided by a rater. You will be provided with these components:
1. **Original Question**: The question that was asked by a user about a video
2. **Model Response**: The model's response to that question
3. **Scoring Data**: The rater's rationales and numerical scores for the model's response
4. **Scoring Criteria**: The standards and criteria that should guide the scoring

Important: You only need to evaluate the scoring data. The other components are provided for your reference only.

Guidelines for Evaluation
Verify the accuracy of the scoring data by evaluating the following aspects:
- **Accuracy of Interpretation**: Do scoring rationales not misrepresent or misunderstand the original meaning of the model's response?
- **Alignment with Scoring Criteria**: Does the scoring data properly follow the established criteria? Specifically:
    - Are the evaluation criteria cited in the rationales consistent with the provided scoring criteria?
    - Based on the situations described in the rationales, are the final scores reasonable and justified according to the scoring criteria?

Here are some points to note:
1. Note that any terms like 'default value' should not appear in the rationale, as they indicate a lack of specific justification for the score.
2. The scoring rationales include descriptions of the video content written by the rater. **Do not** verify them.

Output JSON Format
{
    "explanation": "A brief justification of why the evaluation passed or failed.",
    "verdict": "Pass/Fail",
}

User Prompt for Judging

Reference Information
Scoring Criteria
There are five dimensions for scoring, each scored on a scale of 0 to 3. The detailed criteria for

each score level are as follows:
{Criteria}

Original Question
{Question}

The Model's Response
{Response}

Scoring Data To Be Verified
{Rationales and Final Scores}

B.4 Prompts for Hallucination Detection

Here are the system prompt and user prompt for detection hallucination of generated scoring data using Gemini 3 Flash.

System Prompt for Hallucination Detection

You are a helpful validator. Your task is to compare the provided video against a given scoring rationale for a model's response.
Determine if the rationale contains hallucinations. A hallucination is defined as any claim that is not supported by the video.

Please note that:
- You only need to verify the scoring rationale, not the model's response.
- The scoring rationale may incorporate the rater's in-depth comprehension of the video content, including characters' mental states or metaphorical cues. If you cannot definitively prove it is erroneous, do not classify it as a hallucination.

Output JSON Format {
"evidence": "A brief justification derived from a comparative analysis of the video and the rationale",
"is_hallucinated": "Whether the rationale contains hallucinations (true/false)",
}

User Prompt for Hallucination Detection

A Model's Response
{Response}

Scoring Rationale
{Rationales}

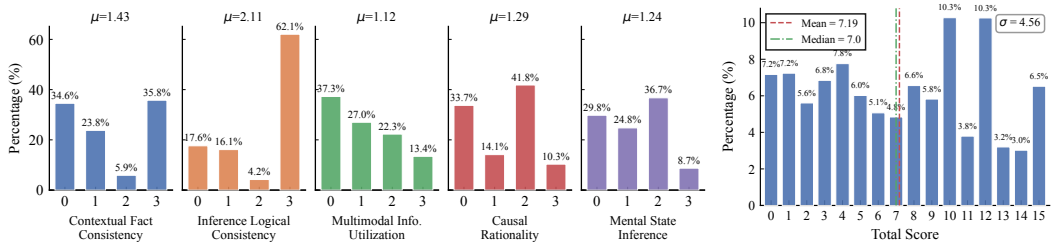
B.5 Statistical Analysis

SocialPrism-Data. Figure 5 details the score distributions of SocialPrism-Data. The five dimensions show diverse patterns: ILC skews high (μ=2.11, σ=1.21, with 62.1% scoring 3), while MIU skews low (μ=1.12, σ=1.06). Total scores (μ=7.19, σ=4.56) cover a broad range of quality levels, with no single bin exceeding 10.3% of the dataset.

SocialPrism-Bench. As shown in Figure 6, our benchmark of 339 human-validated samples exhibits distinct score distributions across dimensions. ILC receives predominantly high scores

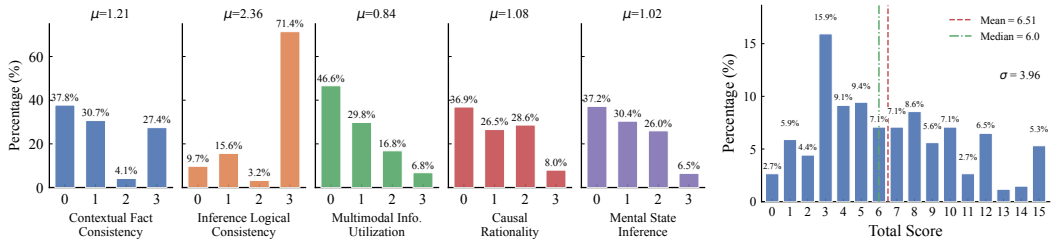
(μ=2.36, σ=1.06, with 71.4% scoring 3), whereas MIU receives lower scores (μ=0.84, σ=0.94). Total scores (μ=6.51, σ=3.96) span the full 0–15 range, with the most frequent score (3) accounting for 15.9% of the benchmark, reflecting broad coverage across response quality levels.

VSR-8K. Questions average 12.2 words in length (σ=4.0). The correct options are slightly longer than the distractors (9.3 vs. 7.8 words on average), as they tend to require more precise descriptions.



(a) Score distribution across dimensions. (b) The distribution of total scores.

Figure 5: Score distributions of the SocialPrism-Data. The five dimensions exhibit different score distributions, while total scores span a broad range of response quality.



(a) Score distribution across dimensions. (b) The distribution of total scores.

Figure 6: Score distributions of the SocialPrism-Bench.

C Details of Gen-SocialPrism-7B

For both training and inference, the video input is up to 16 frames with a maximum of 150,528 pixels per frame, and audio is processed by default.

The system prompt used for training and inference when doing point-wise scoring task is: Here is the system prompt for Gen-SocialPrism-7B when doing point-wise scoring tasks. The placeholders "Scoring Criteria" and "Output format" are the same as that used in Section B.2.

System Prompt for Gen-SocialPrism-7B When Doing Point-wise Scoring Tasks

Task
You will act as an impartial judge. Your task is to meticulously evaluate the quality of a model's responses to a specific question about a video consisting of rich social interactions. The response consists of two parts: a "thinking process" and a "final answer". Here are some guidelines for evaluation:
- You are provided with predefined evaluation dimensions. For each dimension, you need to thoroughly evaluate the quality of the model's response within that dimension, and then provide reasons for your scoring and the corresponding score (0 to 3).
- All your evaluation must be grounded in the provided information about the video.
- Do not let the length or writing style of the response influence your evaluation.
- Your evaluation of the "Inference Logical Consistency" dimension must be based exclusively on the model's response. For other dimensions, you must evaluate the model's response in

conjunction with the video context.

Predefined Evaluation Dimensions
{Scoring Criteria}

Inputs
You will receive a video, the question, and the model's response.

Output Format
{Output format}

Here is the system prompt for Gen-SocialPrism-7B when doing pair-wise scoring tasks.

System Prompt for Gen-SocialPrism-7B When Doing Pair-wise Scoring Tasks

Task
You will act as an impartial judge. Your task is to meticulously evaluate and compare the quality of two models' responses (Model A and Model B) to a specific question about a video consisting of rich social interactions. Each response consists of two parts: a "thinking process" and a "final answer". Here are some guidelines for evaluation:
- You are provided with key evaluation criteria. You need to thoroughly compare the models' responses based on these criteria, weigh their relative strengths and weaknesses, and decide which model provides the better overall response.
- All your evaluation must be grounded in the provided information about the video.
- Do not let the length or writing style of the response influence your evaluation. Favor substance, accuracy, and depth over mere verbosity.
- Your evaluation of the "Inference Logical Consistency" criterion must be based exclusively on the models' responses. For other criteria, you must evaluate the models' responses in conjunction with the video context.

Evaluation Criteria
When comparing the two models, consider their performance across the following dimensions:

Contextual Fact Consistency: Evaluate whether the objective information described by the model is entirely faithful to the video content. Penalize models that contain factual errors, hallucinations, or critical errors on key facts that cause the analysis to be based on a flawed premise.
Inference Logical Consistency: Evaluate whether the model's response is free from internal logical contradictions. This judges the structure of the argument, not its plausibility or factual accuracy. You need to evaluate the internal logical consistency of the thinking process as well as the logical consistency between the thinking process and the final answer.
Multimodal Information Utilization: Evaluate whether the model effectively integrates information from both visual and auditory (dialogue) modalities. Reward models that weave together visual details (e.g., character expressions, actions, environment) and dialogue to derive insights that would be impossible to obtain from either modality alone.
Causal Rationality: Evaluate the plausibility and depth of the model's explanations for why events, behaviors, or mental states occur. It judges the soundness of the connective "because" link in the model's reasoning. Reward comprehensive and logically rigorous explanations over superficial or far-fetched ones.
Mental State Inference: Evaluate the ability to accurately and insightfully infer mental states. This includes identifying observable and straightforward mental states (e.g., emotions shown on the face, expressed beliefs), as well as uncovering deeper, hidden motivations, temperaments, desires, beliefs, and intentions based on the evidence.

Inputs

You will receive a video, the question, Model A's response, and Model B's response.

Output Format
Please output your evaluation in the following JSON format:
{
"rationale": "A detailed comparison of Model A and Model B across the evaluation criteria. Highlight specific strengths, weaknesses, and errors in both models, and provide a clear justification for why one model is overall better than the other.",
"winner": "The winning model. Must be one of: 'Model A' or 'Model B'."
}

D Details of Experiments

D.1 Details of "+ Gen-SocialPrism-7B" Setting

The overall reward fuction is $r = r_{\texttt{acc\&sp}} + 0.5 \cdot r_{\texttt{fmt}}$ $r_{\texttt{acc\&sp}}$ is a new reward integrating accuracy and the reward signal from Gen-SocialPrism-7B:

$$r_{\texttt{acc\&sp}} = \begin{cases} \alpha \cdot r_{\texttt{sp}} + (1-\alpha) & \text{if correct} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $r_{\texttt{sp}} = \frac{1}{15} \cdot \mathbb{I}[s^{\texttt{CFC}} > 0 \& s^{\texttt{ILC}} > 0] \cdot \sum_{c \in \mathcal{C}} s^c$ denotes the normalized total score from Gen-SocialPrism-7B and $\alpha = 0.2$. This formulation requires the model to perform reasoning without significant factual hallucinations or logical flaws. Furthermore, it ensures that among correct responses, those exhibiting higher social reasoning quality receive proportionally larger rewards, enabling finer-grained optimization beyond binary correctness.

D.2 Hyperparameter Sensitivity Analysis under "+ Gen-SocialPrism-7B"

We evaluate the sensitivity to α under the "+ Gen-SocialPrism-7B" setting using values of 0.15, 0.20, and 0.30 while keeping all other settings fixed. As illustrated in Figure 7, the default setting, $\alpha = 0.20$, achieves the highest accuracy on EmoBench-M SD and MPDER, as well as the longer-video subset of OmniVideoBench, whereas $\alpha = 0.15$ performs better on PEA, WorldSense, and the shorter-video subset. Both settings outperform the RPO and GRPO baselines across all eight metrics. These results reveal task-dependent trade-offs and support $\alpha = 0.20$ as a reasonable default rather than a uniformly optimal choice.

D.3 Details of Reward-Guided Voting

For each benchmark b, let $\mathcal{C}_b \subseteq \mathcal{C}$ denote the set of evaluation dimensions used to construct the test-time reward signal. For social reasoning benchmarks, $\mathcal{C}_b = \mathcal{C}$; for omni-modal benchmarks, we exclude the MSI dimension since these benchmarks typically do not require explicit mental state inference.

Given the i-th candidate response, we first compute its normalized aggregated reward score as

$$\rho_i = \frac{1}{3|\mathcal{C}_b|} \sum_{c \in \mathcal{C}_b} s_i^c \tag{8}$$

where $s_i^c \in \{0, 1, 2, 3\}$ denotes the score assigned to the i-th response on dimension c. The factor $3|\mathcal{C}_b|$ normalizes the aggregated score to the range $[0, 1]$.

During sampling, each candidate response votes for its parsed answer a_i with a reward-guided hinge weight:

$$w_i = 1 + \lambda \max(0, \rho_i - \tau), \tag{9}$$

where τ is the reward threshold and λ controls the strength of reward amplification. The final prediction is obtained by summing the weights of all responses that yield the same parsed answer:

$$\hat{a} = \arg\max_a \sum_{i: a_i = a} w_i. \tag{10}$$

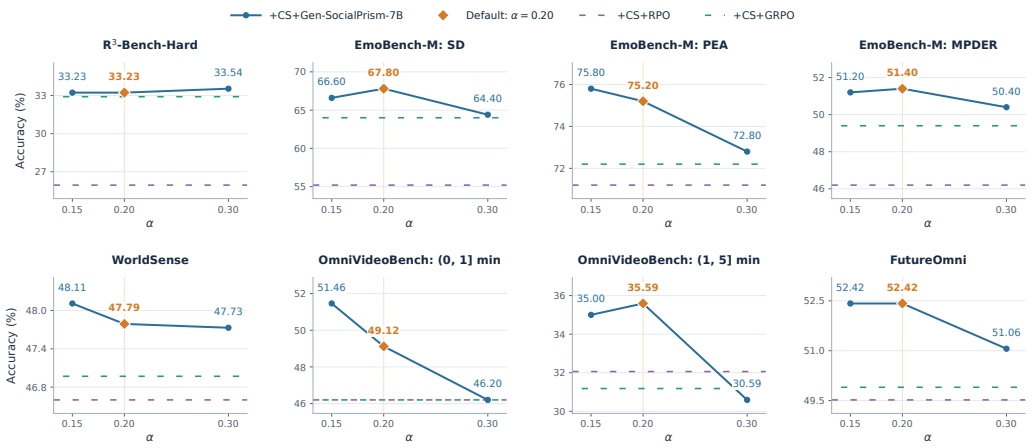


Figure 7: Sensitivity to α under the "+ Gen-SocialPrism-7B" setting across eight benchmark subsets, with all other settings held fixed. Orange diamonds mark the default value $\alpha = 0.20$. Purple and green horizontal dashed lines indicate the "+ CS + RPO" and "+ CS + GRPO" baselines, respectively. Accuracy (%) is reported using an independent y-axis range for each subset.

When all $w_i = 1$, this procedure reduces to standard majority-of-N voting.

D.4 Prompts for GRPO Training

System Prompt for GRPO Training

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

User Prompt for GRPO Training

{question}
{options_text}

D.5 Evaluation Configurations of Post-trained Models

For R^3-Bench-Hard and EmoBench-M, we sample up to 16 frames. For WorldSense and OmniVideoBench "(1,5] min", we sample up to 64 frames; and for FutureOmni and OmniVideoBench "(0,1] min", we sample up to 32 frames.

Unlike the training QA dataset, VSR-8K, EmoBench-M lacks explicit option letters, which introduces a potential domain shift. To mitigate this, when evaluating the post-trained models, we apply a semantically lossless transformation to restructure it into a standard multiple-choice format with option letters, preserving the fundamental nature of the task.

D.6 Details of Test-Time Scenarios

D.6.1 Evaluation Configuration

For R^3-Bench-Hard and EmoBench-M, we sample up to 16 frames. For FutureOmni, we sample up to 32 Frames. For WorldSense and OmniVideoBench, we sample up to 64 frames.

D.6.2 PROMPTS

In the test-time scenarios, we use *You are a helpful assistant.* as the system prompt. User prompts used in each benchmark are as follows. We appropriately modify the official evaluation prompts (some of which originally only require the final answer) to require the output of the thinking process.

User Prompt for Inference for R^3-Bench-Hard

These are the frames of a video and the corresponding audio. Select the best answer to the following multiple-choice question based on the video. Output only the letter (A, B, C, D, or E) of the correct option as your choice first, and then provide your thinking process.
Question: {question}
{options_text}

User Prompt for Sampling for R^3-Bench-Hard

These are the frames of a video and the corresponding audio. Select the best answer to the following multiple-choice question based on the video. Output only the letter (A, B, C, D, or E) of the correct option as your choice first, and then provide your thinking process.
Question: {question}
{options_text}
Answer:

User Prompt for Inference and Sampling for EmoBench-M SD

The person in video says: <Subtitle Placeholder>. Does this statement express sarcasm? Provide a detailed thinking process and choose one of the following labels as your final answer: true, false. Respond in the format: {'thinking_process': 'your thinking process', 'answer': 'label'}.

User Prompt for Inference and Sampling for EmoBench-M PEA

The person in video says: <Subtitle Placeholder>. Determine the emotion conveyed. Provide a detailed thinking process and choose one of the following labels as your final answer: neutral, negative, positive. Respond in the format: {'thinking_process': 'your thinking process', 'emotion': 'label'}.

User Prompt for Inference and Sampling for EmoBench-M MPDER

The person in video says: <Subtitle Placeholder>. Please watch the provided video and determine the emotion it conveys. Provide a detailed thinking process and choose one of the following labels as your final answer: neutral, surprise, fear, sadness, joy, disgust, anger. Respond in the format: {'thinking_process': 'your thinking process', 'emotion': 'label'}.

User Prompt for Inference and Sampling for FutureOmni

These are the frames of a video and the corresponding audio.
Select the best answer to the following multiple-choice question based on the video. Output only the letter (A, B, C, D, E, F) of the correct option as your choice first, and then provide your thinking process.

Question: {question}
Options: {options_text}

User Prompt for Inference and Sampling for WorldSense

Carefully watch this video and pay attention to every detail. Based on your observations, select the best option that accurately addresses the question.

These are the frames of a video and the corresponding audio. **First, provide a step-by-step thought process based on the visual and audio evidence.** Then, conclude with the single best answer letter (A., B., C., or D.). You must choose the most likely option; do not answer "None" or claim there is no answer.
[Question]:
{question}
{options_text}

User Prompt for Inference and Sampling for OmniVideoBench

Instruction: You are given a video. Based on the content of the video, answer the following question:
Question: {question}
Options: {options_text}
Task:
Note that you should first reason step by step, and then you should give your final choice in A, B, C, or D.
Your answer format should be as follows:
Step X: [Reasoning step X]
The final choice is:
\bbox {{Answer with the option's letter directly(A, B, C, or D).}}.