AFFTOM: BENCHMARKING RECURSIVE AFFECTIVE THEORY OF MIND IN DYNAMIC MULTI-PARTY INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding emotion in realistic social interactions requires more than recognizing facial expressions or assigning affective labels. The same event can elicit different emotions depending on what each participant knows, what they believe others know or feel, and how these nested mental states evolve over time. Existing benchmarks largely study higher-order epistemic Theory of Mind (ToM) and multimodal emotion reasoning in isolation, leaving recursive affective reasoning under asymmetric and dynamically changing information underexplored. Therefore, we introduce AffToM, a benchmark for affective ToM with 1,922 test pairs and 19,759 training pairs. AffToM evaluates Individual Affective Reasoning (IAR), which infers emotions from events and socio-cognitive context, and Multi-party Affective Reasoning (MPAR), which recursively tracks participants' nested beliefs and emotions as new evidence emerges. We develop an event-driven multi-agent generation pipeline grounded in dynamic character interaction graphs that explicitly capture event sequences, interpersonal relations, and mental-state updates. Furthermore, we propose AffToM-R1, a two-stage post-training framework combining supervised fine-tuning with perspective-aware rubric-guided reinforcement learning. Experiments show that MPAR remains challenging, with even the strongest evaluated MLLM achieving less than 60% accuracy, while AffToM-R1 improves MPAR accuracy by 13.38 percentage points over its base model.

1 INTRODUCTION

Recent Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in visually grounded affective perception, including facial expression recognition, emotion description, and action understanding (Lian et al., 2025; Cheng et al., 2024; Li et al., 2025a). However, recognizing an apparent emotion is fundamentally different from understanding it within the person's broader cognitive and social context, as the same expression, such as a smile, may signal happiness, relief, embarrassment, concealment, or an attempt to reassure others. Genuine emotional understanding requires reasoning beyond perceptual appearance to determine what an individual knows, expects, and believes, how they interpret the situation, and how they understand the mental states of others.

This capability is commonly referred to as *affective Theory of Mind* (ToM), namely the ability to infer emotional states from the latent cognitive representations that shape them. In realistic social interactions, emotions are not direct and universal responses to observable events but emerge from how individuals appraise those events under their beliefs, goals, expectations, and social relationships (Premack & Woodruff, 1978; Lillard, 1993; Qu et al., 2015). Consequently, the same event may evoke different emotions in different participants, while perceptually similar reactions may reflect fundamentally different mental states. The challenge becomes greater in multi-party interactions, where an individual's emotion may depend not only on their own beliefs but also on what they believe another participant knows, intends, or feels about someone else. For instance, a person may feel relieved because they believe a misunderstanding remains unnoticed, embarrassed because they infer that another participant has recognized their concealed intention, or guilty because they realize that one participant now understands that a third participant feels betrayed. In fact, to evaluate machine ToM, a growing body of benchmarks has been proposed (Chen et al., 2024; Wu et al., 2023; Xu et al., 2024; Jin et al., 2024; Shi et al., 2025).

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.

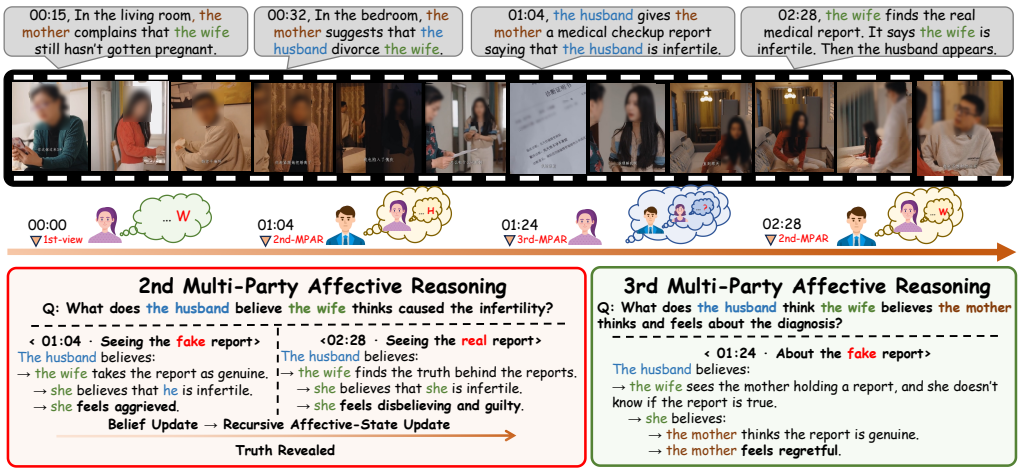


Figure 1: An AffToM example showing multi-party affective reasoning and emotion shifts under information asymmetry.

Despite this progress, existing benchmarks leave two closely related challenges unresolved. First, existing higher-order ToM and affective reasoning capabilities are largely studied in isolation. Higher-order benchmarks focus on recursively nested epistemic states, whereas affective benchmarks typically predict each participant's emotion independently, failing to capture cross-participant affective dependencies in which one person's emotion is shaped by what they believe others know, believe, or feel. Second, current benchmarks provide only limited control over information asymmetry, recursion depth, and temporal belief updates, making it difficult to distinguish genuine higher-order affective reasoning from reliance on superficial perception. These limitations motivate the study of **Multi-party Affective Reasoning (MPAR)**, which evaluates affective states whose correct interpretation depends on recursively nested mental representations distributed across multiple participants. Within an ongoing social interaction, an MPAR question asks how another person believes the target character feels before the character's emotional response becomes observable. As illustrated in Figure 1, the husband's affective inference changes as his information state evolves. After the wife observes a fabricated medical report, he infers that she mistakenly believes he is infertile and interprets her affective state as feeling wronged. When he later observes her discovering the authentic report, he updates his representation of her belief and instead infers disbelief and guilt. Although the participants and broader situation remain largely unchanged, the correct affective inference changes because the husband's representation of what the wife knows and believes has evolved. Therefore, we introduce AffToM, a video-based benchmark for affective ToM that evaluates latent emotion inference from characters' evolving cognitive representations. It contains 1,922 QA pairs from 258 videos for benchmark evaluation, complemented by 19,759 QA pairs from 2,291 videos for model training. AffToM evaluates two complementary capabilities: **Individual Affective Reasoning (IAR)**, which requires models to infer an individual's affective state by integrating observable events with situational context, and **MPAR**, which further requires models to recursively track how participants represent one another's beliefs and emotions and use these nested representations to infer affective states. We develop an event-driven multi-agent question-generation pipeline grounded in dynamic character interaction graphs that explicitly represent event chronology, interpersonal relations, and mental-state transitions. These structured graphs enable controlled generation of questions with different participant configurations, information asymmetries, temporal update depths, and levels of mental-state recursion. Role-specialized agents generate IAR and MPAR questions from the graphs, after which the resulting samples are refined through an iterative dual-criteria evolution process that jointly improves logical validity and distractor discriminability, followed by human validation. These findings motivate us to examine whether targeted training can improve belief-conditioned emotion inference from others' perspectives. We introduce AffToM-R1, a two-stage post-training framework that applies supervised fine-tuning (SFT) followed by perspective-aware rubric-guided reinforcement learning. Comprehensive evaluations of state-of-the-art MLLMs under both multiple-choice and open-ended settings show that MPAR remains particularly challenging, with even the strongest evaluated model achieving less than 60% accuracy. Meanwhile, AffToM-R1 improves MPAR accuracy by 13 percentage points over its Qwen3.5-4B base model.

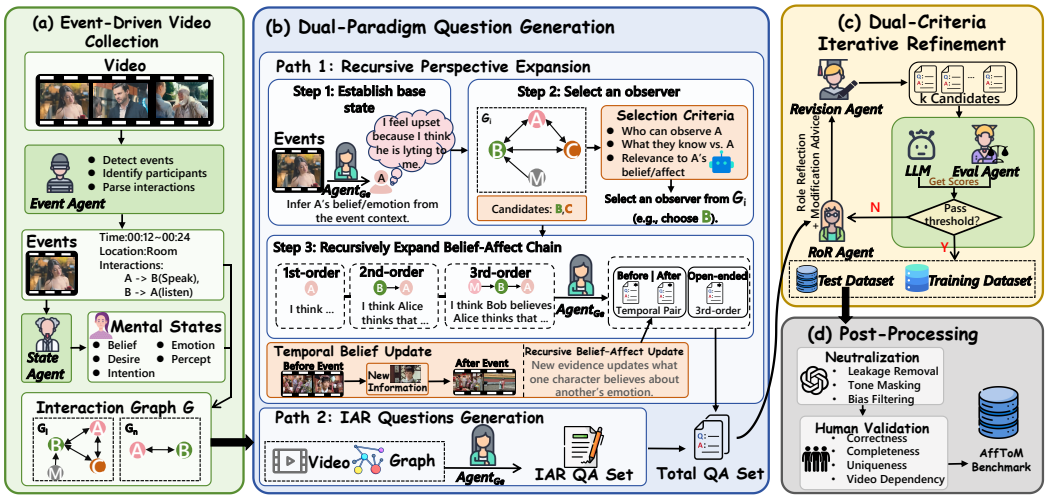


Figure 2: Overview of the dataset construction pipeline, including event-driven video collection, question generation, data evolution, and manual evaluation.

2 Related Work

Theory of Mind. Theory of Mind (ToM) refers to the capacity to infer others' mental states, including beliefs, intentions, and emotions (Premack & Woodruff, 1978; Decety & Jackson, 2004; Qu et al., 2015). Text-based benchmarks evaluate belief tracking, false-belief reasoning, and recursive perspective-taking (Le et al., 2019; Kim et al., 2023; Wu et al., 2023; Xu et al., 2024; Chen et al., 2024), while recent multimodal benchmarks extend ToM to visually grounded and interactive settings. MMToM-QA and MuMA-ToM study beliefs and goals in multi-agent environments (Jin et al., 2024; Shi et al., 2025), whereas MOMENTS, Social-IQ 2.0, and IntentQA evaluate broader social reasoning abilities from real-world videos (Villa-Cueva et al., 2025; Wilf et al., 2023; Li et al., 2023). More recent work further explores mental-physical causal reasoning and multimodal emotion reasoning (Niu et al., 2026; Luo et al., 2026). Despite this progress, emotion inference through nested mental states and its evolution remains underexplored.

Multimodal Question Answering. Multimodal QA has evolved from recognizing visual content toward integrating temporally distributed evidence and reasoning about event relations. Video benchmarks such as PerceptionTest and HourVideo evaluate temporal, causal, predictive, and long-context reasoning (Pătrăucean et al., 2023; Chandrasegaran et al., 2024). In the affective domain, early benchmarks primarily evaluate multimodal sentiment and emotion recognition with predefined labels (Bagher Zadeh et al., 2018; Poria et al., 2019; Lian et al., 2023). These advances move beyond emotion classification toward contextual affective understanding, but largely stop short of reasoning over recursively dependent affective states across multiple participants.

3 AffToM

In this section, we detail the systematic construction of the AffToM benchmark. We introduce a multi-stage data generation pipeline driven by a collaborative multi-agent framework, with overall structure shown in Figure 2. This systematic approach sequentially consists of event-driven video curation, dual-paradigm question generation enhanced by a social interaction graph, dual-criteria-guided data evolution, and comprehensive post-processing alongside expert validation.

3.1 Event-Driven Video Collection

We collect 1,504 raw video candidates from public video-sharing platforms, including Douyin, TikTok, and YouTube. Collection details are provided in Appendix B.1. These videos span diverse social scenarios, thereby providing rich multimodal signals for modeling complex human behaviors. To systematically parse the semantic content of these candidate videos, we employ Seed-1.8 (Seed,

2026a) to extract fine-grained event metadata. For video v, the event extraction is defined as:

$$\mathcal{E} = \text{Agent}_{\text{eve}}(v), \tag{1}$$

where $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$ denotes the set of events extracted from the video. Each event e_i consists of its spatiotemporal grounding—comprising the timestamp, location, and involved individuals. To capture multi-party social dynamics for affective ToM reasoning, we retain videos with multiple characters connected by explicit directed interaction pairs. This filtering yields 308 videos, for which we extract character-level cognitive and affective attributes:

$$\mathcal{S} = \text{Agent}_{\text{state}}(v, \mathcal{E}). \tag{2}$$

Here, $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ denotes the set of individual-level socio-cognitive states, encompassing identity traits, interpersonal relationships, and dynamically evolving mental states of the characters throughout the video, where s_i corresponds to the i-th character c_i. Using events $\mathcal{E}$ and individual-level information $\mathcal{S}$, we construct a dynamic character interaction graph $G = \{G_1, G_2, \ldots, G_N\}$, where nodes represent the characters and edges encode their evolving behavioral relationships.

3.2 Dual-Paradigm Question Generation

After constructing the dynamic character interaction graph G, we design two complementary question generation paths to construct affective ToM reasoning tasks.

Recursive Perspective Expansion. The first paradigm generates Multi-party Affective Reasoning (MPAR) questions through recursive perspective expansion. For each extracted event, we establish a base-level affective state or belief for a target individual, grounded in the event context:

$$\mathcal{Q}^{(1)} = \Phi_{\text{init}}(\mathcal{E}, v), \tag{3}$$

$\mathcal{Q}^{(1)} = \{q_1^1, q_2^1, \ldots, q_N^1\}$ serves as the foundational belief questions from which more complex reasoning structures are derived. To construct MPAR questions, we progressively introduce additional individuals from the interaction graph G and recursively integrate their respective perspectives into the reasoning process. Second-order questions introduce a second individual and require reasoning about one person's belief concerning another's mental or emotional state. Third and higher orders add further parties and their perspectives. Rather than merely nesting beliefs, the questions capture how attitudes, expectations, and misunderstandings shape emotions within the shared interaction, evaluating dynamic, interaction-driven affective reasoning beyond isolated emotion recognition. We formally define this recursive generation of higher-order cognitive states as:

$$(\mathcal{Q}^{(i+1)}, \mathcal{A}^{(i+1)}) = \Phi_{\text{recur}}(\mathcal{Q}^{(i)}, G_i, v), \tag{4}$$

where $\mathcal{Q}^{(i)}$ denotes the i-th order question and $\mathcal{A}^{(i+1)}$ denotes the corresponding answer. Φ_{recur} is the recursive integration function that expands the current reasoning structure by incorporating the perspective of a newly introduced character c_k, using the corresponding event and character information encoded in G_i, together with the video v.

To quantify the behavioral effect of the interaction, we formulate some higher-order queries as temporal pairs spanning across the event timeline. Specifically, we generate coupled questions querying the nested affective states at both a pre-event timestamp t_{pre} and a post-event timestamp t_{post}. This contrasting design forces models to explicitly trace the temporal evolution of emotional states and measure the causal impact of the specific event on resolving or exacerbating interpersonal misunderstandings. Upon deriving the definitive question-answer pairs $(\mathcal{Q}, \mathcal{A})$, we synthesize a set of semantically plausible distractors to construct the final multiple-choice questions (MCQs). Furthermore, to complement this standard discriminative evaluation, we introduce open-ended MPAR questions specifically designed to assess deeply nested higher-order affective reasoning. An event-driven MPAR generation example is provided in Appendix F.2.

Individual Affective Reasoning Generation. Complementing multi-party relational reasoning, the second paradigm is designed to formulate individual affective reasoning (IAR) questions that probe the internal states and latent characteristics of individual characters within emotionally rich social contexts. Specifically, this track covers six dimensions of affective and behavioral understanding, including Preference Prediction, Integrated Mental-State Reasoning, Strategic Perspective-Taking, Hidden Intention, Personality Reasoning, and Counterfactual Reasoning. For this type, we generate the question through a question generation agent with video v and the graph G as inputs, guided by a curated set of few-shot exemplars derived from human-validated seed questions.

3.3 DUAL-CRITERIA ITERATIVE REFINEMENT

To improve cognitive depth and linguistic quality, our iterative pipeline combines role-guided reflection with dual-criteria evaluation. The Role Reflection Agent ($\text{Agent}_{\text{RoR}}$) examines each generated question from the event's different roles, identifying improvements, ambiguities, and alternative viewpoints grounded in the interaction structure. Using this feedback, the Revision Agent ($\text{Agent}_{\text{rev}}$) refines the initial drafts into a diverse candidate pool $\mathcal{Q}_{\text{cand}}$. We then select high-quality questions from $\mathcal{Q}_{\text{cand}}$ using two criteria: *Comparative Rationality* and *Distractor Unbiasedness*.

Criterion 1: Comparative Rationality (S_{rank}). Most existing data generation pipelines adopt scoring-based selection strategies to identify high-quality candidates from a pool of generated samples (Li et al., 2025b; Agarwal et al., 2025). However, relying on standard absolute scoring or uncalibrated list-wise ranking often introduces severe vulnerabilities, such as verbosity, self-enhancement, and position biases (Zheng et al., 2023). To rigorously isolate the highest-quality questions and circumvent these biases, we employ a robust pairwise comparison mechanism benchmarked against human expertise. Specifically, we prompt an evaluation agent (i.e., $\text{Agent}_{\text{eval}}$) to conduct blind pairwise evaluations between a generated candidate $q_c \in \mathcal{Q}_{\text{cand}}$ and a sampled human-authored reference question $q_h \in \mathcal{Q}_{\text{human}}$ from the seed pool. $\text{Agent}_{\text{eval}}$ judges the candidate superior, equivalent, or inferior to the human reference based on logical coherence, contextual alignment, and affective depth. The comparative rationality score is the expected win rate over the human references:

$$S_{\text{rank}}(q_c) = \frac{1}{|\mathcal{Q}_{\text{human}}|} \sum_{q_h \in \mathcal{Q}_{\text{human}}} \Big(\mathbb{I}(q_c \succ q_h) + \tfrac{1}{2}\mathbb{I}(q_c \sim q_h) \Big). \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $q_c \succ q_h$ denotes that the candidate outperforms the human reference, and $q_c \sim q_h$ denotes a tie in quality. This pairwise formulation, bounded between 0 and 1, mitigates list-wise position bias and ensures that a high score reflects a quality level authentically comparable to, or explicitly exceeding, expert human annotations.

Criterion 2: Distractor Unbiasedness (S_{bias}). In multimodal benchmarks, models often exploit linguistic priors to guess the correct answer without processing the visual input. To ensure our distractors (i.e., incorrect options) are plausible and effectively penalize blind guessing, we evaluate the candidate options $\mathcal{O} = \{A, B, C, D\}$ under a video-blind setting. Given only the textual context, we estimate the prior probability distribution over the options using an LLM. The LLM produces output logits for each candidate option, which are converted into a probability distribution $P(o \mid \text{text-only})$ via a softmax function. An unbiased question should compel the text-only model into a state of maximum uncertainty, yielding a uniform distribution across all options. We formally quantify this uncertainty using Normalized Shannon Entropy:

$$S_{\text{bias}}(q_c) = -\frac{1}{\log |\mathcal{O}|} \sum_{o \in \mathcal{O}} P(o \mid \text{text-only}) \log P(o \mid \text{text-only}). \tag{6}$$

If the model is highly confident in a single option without seeing the video, S_{bias} yields a low score, indicating that the distractors are ineffective or the question is textually biased.

Iterative Evolution and Termination. We combine the two criteria into a unified score for iterative candidate refinement:

$$S_{\text{total}}(q_c) = \lambda S_{\text{rank}}(q_c) + (1 - \lambda) S_{\text{bias}}(q_c), \tag{7}$$

where λ controls the trade-off between question quality and distractor robustness. If a generated candidate meets a predefined acceptance threshold τ, it is accepted as a high-quality question. Otherwise, the generation process enters the next evolution cycle. This iterative reflection–revision–ranking process continues until a candidate question satisfies the criterion. To prevent unbounded iterations, we impose a predefined maximum number of refinement rounds T. The dynamic selection and routing process for a candidate at iteration t is defined as:

$$q^* = \begin{cases} q_c^{(t)} & \text{if } S_{\text{total}}\left(q_c^{(t)}\right) \geq \tau, \\ \Phi_{\text{evolve}}\left(q_c^{(t)}\right) & \text{if } S_{\text{total}}\left(q_c^{(t)}\right) < \tau \text{ and } t < T_{\max}, \\ \emptyset & \text{if } t \geq T_{\max}. \end{cases} \tag{8}$$

where $\Phi_{\text{evolve}}(q_c^{(t)}) = \text{Agent}_{\text{rev}}\big(\text{Agent}_{\text{RoR}}(G_i, q_c^{(t)})\big)$ encapsulates the role-guided reflection and revision operations. τ is a predefined acceptance threshold, and $T_{\max}$ represents the maximum allowable number of refinement cycles.

3.4 Post-Processing and Human Validation

The preceding generation pipeline produces 19,759 QA samples from 2,291 videos for model training. The test set undergoes two additional steps: linguistic post-processing and human validation. First, to reduce linguistic shortcuts, we apply linguistic neutralization to the questions and options retained after iterative refinement. We remove explicit emotional polarity and soften overly deterministic phrasing, for example by rewriting "react negatively" as "show a particular emotional reaction". This encourages models to reason about multimodal context rather than rely on textual cues. We manually validate the questions against the source videos (see validation details in Appendix C.1). Through this multi-stage filtering and validation process, we obtain a benchmark containing 1922 high-quality questions grounded in 258 videos. Detailed statistics are provided in Appendix B.1.

4 AffToM-R1

In this section, we present AffToM-R1. As illustrated in Figure 3, SFT trains the model on structured reasoning trajectories, while Perspective-Aware Faithfulness Reasoning Optimization (PAFRO) further optimizes it with perspective-aware rubric-based rewards. We construct a 20k-sample training pool and split it into $\mathcal{D}_{\mathrm{SFT}}$ for SFT and $\mathcal{D}_{\mathrm{RL}}$ for PAFRO based on difficulty estimated from the pretrained backbone. Further training details are provided in Appendix D.1.

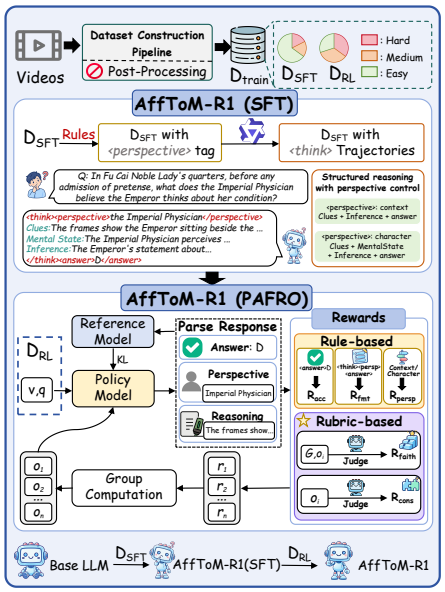


Figure 3: Overview of AffToM-R1.

4.1 Supervised Fine-Tuning

We first perform supervised fine-tuning (SFT) on perspective-structured reasoning trajectories that connect video evidence to the target answer. Each target response consists of a reasoning block enclosed by `<think></think>` and a final answer enclosed by `<answer></answer>` (see Appendix D.2 for more details). The reasoning block begins with a `<perspective></perspective>` field that explicitly identifies the viewpoint from which the question is resolved to guide later reasoning.

Perspective-structured reasoning. We adopt two reasoning structures depending on whether the question requires character-specific mental-state inference. For the *character-centered* structure, the trajectory follows *Clues* $\rightarrow$ *Mental State* $\rightarrow$ *Inference*, explicitly interpreting the observed evidence through the target character's information and mental state before deriving the answer. For the *context-centered* structure, the trajectory follows *Clues* $\rightarrow$ *Inference* when explicit character-level mental-state modeling is unnecessary. All MPAR questions adopt the *character-centered* structure, while the remaining questions are assigned an appropriate structure with LLM assistance. We generate candidate trajectories accordingly and retain only those whose final answers match the reference answers, yielding $\mathcal{D}_{\mathrm{SFT}}$ for SFT.

4.2 Perspective-Aware Faithfulness Reasoning Optimization

PAFRO further optimizes the SFT-trained model using rubric-based GRPO, as illustrated in Figure 3. Details of GRPO are provided in Appendix D.3.

Reward design. Answer correctness alone does not reveal whether the model adopts the appropriate viewpoint or supports its conclusion with video evidence. We therefore combine rule-based checks of answer correctness (R_{acc}), response format (R_{fmt}), and perspective alignment (R_{persp}) with two reasoning rubrics. Mental-state faithfulness (R_{faith}) checks the response's claims about beliefs, intentions, and emotions by comparing them with the dynamic character interaction graph G provided as reference, using the relevant event or whole-video scope. For *character-centered* responses, this

Table 1: Performance comparison of different MLLMs. Human performance is evaluated on 250 questions randomly sampled from AffToM (see Appendix C.2 for more details).

Models	Params	Multiple-Choice Question								Open-ended		
		PP	IMSR	SPT	HI	PR	CR	MPAR	Avg.	EC	FC	Avg.
Open-Source MLLMs												
Ministral-3-3B	3B	31.02	39.81	43.15	27.49	38.89	33.33	30.07	33.26	1.64	1.48	1.56
Gemma-3-4B	4B	40.11	52.43	43.15	39.77	42.78	37.85	29.67	37.43	1.62	1.26	1.44
LLaVA-OneVision-1.5	4B	47.59	63.11	60.96	47.95	45.00	48.59	**39.34**	46.87	1.82	1.51	1.67
Qwen3-VL-4B-Instruct	4B	50.00	**63.41**	51.37	53.80	44.44	46.89	36.42	45.50	**1.88**	**1.57**	**1.73**
Qwen3.5-4B	4B	**52.13**	61.46	**66.44**	**55.56**	**54.44**	**58.19**	37.22	**49.29**	1.77	1.50	1.64
Ministral-3-8B	8B	44.92	50.00	48.63	45.03	48.89	40.68	32.05	40.45	1.79	1.60	1.70
LLaVA-OneVision-Qwen2	7B	40.64	50.00	52.74	47.95	41.11	44.63	33.91	41.00	1.85	1.48	1.67
InternVL3.5-8B	8B	50.27	52.91	58.22	50.29	53.33	54.24	32.32	44.46	1.88	1.60	1.74
MiniCPM-V-2.6	7B	44.39	61.65	61.64	49.71	47.22	53.67	36.03	45.94	1.76	1.58	1.67
Qwen2.5-VL-7B-Instruct	7B	52.41	58.74	58.22	50.29	52.22	53.67	36.03	46.71	1.85	1.58	1.72
LLaVA-OneVision-1.5	8B	44.39	**67.48**	63.01	52.63	56.67	57.63	33.77	47.37	**1.96**	1.61	1.79
MiniCPM-V-4.5	8B	47.59	62.62	67.12	60.23	56.11	55.93	39.60	50.38	1.49	1.31	1.40
Qwen3-VL-8B-Instruct	8B	**59.36**	66.99	58.90	**61.99**	**59.44**	54.80	39.87	51.92	1.92	**1.70**	**1.81**
Qwen3.5-9B	9B	57.22	**67.48**	**68.49**	**61.99**	57.22	**64.97**	**40.40**	**53.51**	1.79	1.52	1.66
Gemma-3-27B	27B	50.27	58.74	45.21	55.56	53.89	45.20	30.07	42.81	1.87	1.45	1.66
InternVL3.5-14B	14B	49.20	54.37	63.70	49.71	48.89	51.41	34.04	44.90	1.82	1.62	1.72
InternVL3.5-38B	38B	52.94	58.25	64.38	46.78	53.33	59.32	35.36	47.26	1.99	1.76	1.88
Qwen2.5-VL-72B-Instruct	72B	54.55	72.33	65.75	61.40	61.67	**66.10**	44.24	55.65	2.22	1.95	2.09
Qwen3-VL-32B-Instruct	32B	61.50	**77.67**	66.44	63.16	**64.44**	64.41	40.79	55.87	2.39	1.96	2.18
Qwen3.5-27B	27B	**67.91**	75.73	**72.60**	**69.01**	63.89	**66.10**	**45.03**	**59.22**	**2.45**	**2.23**	**2.34**
Closed-Source MLLMs												
Seed-2.0-lite	–	65.78	72.82	**71.23**	70.76	62.78	66.10	44.24	58.29	2.02	1.88	1.95
GPT-5.5	–	**71.28**	**76.59**	68.49	**71.93**	**66.67**	**74.01**	**52.05**	**63.56**	**3.06**	**2.99**	**3.03**
Ours												
AffToM-R1 (Qwen3.5-4B)	4B	54.79	73.17	72.60	63.16	63.33	66.67	50.60	59.33	2.17	1.89	2.03
Human Baseline	–	91.30	92.59	90.48	95.00	92.00	91.67	84.55	88.80	–	–	–

rubric assesses claims from the declared character's viewpoint. *context-centered* responses are assessed without a single-character viewpoint. Evidence-answer consistency (R_{cons}) checks whether sampled frames and subtitles support the reasoning and whether that reasoning supports the selected answer. It does not use the reference answer, allowing the reward to distinguish evidential support from answer correctness. The response-level reward is:

$$r_i = \lambda_{\text{acc}} R_{\text{acc}} + \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{persp}} R_{\text{persp}} + \lambda_{\text{faith}} R_{\text{faith}} + \lambda_{\text{cons}} R_{\text{cons}}, \tag{9}$$

where each component is evaluated for response y_i and lies in $[0, 1]$. The λ coefficients set their relative weights. Appendix D.4 provides the scoring rules for all five components.

5 Experiments

5.1 Experimental Settings

Evaluated Models. We evaluate open-source and closed-source MLLMs across model families and scales, as listed in Table 1. For the training study, we compare Qwen3.5-4B, its SFT variant, and AffToM-R1 on AffToM and external benchmarks covering social reasoning and higher-order belief reasoning. Model families and deployment details are described in Appendix D.5.

Evaluation Metrics. For MCQs, we utilize standard accuracy as the primary quantitative metric. For open-ended queries, we further employ GPT-5-mini to score the generated responses on a standard 1-to-5 scale across Emotion Correctness (EC) and Factual Consistency (FC), where 1 indicates poor quality and 5 indicates high quality. **EC** evaluates whether the predicted affective state accurately matches the ground-truth emotion experienced by the target character. **FC** measures the extent

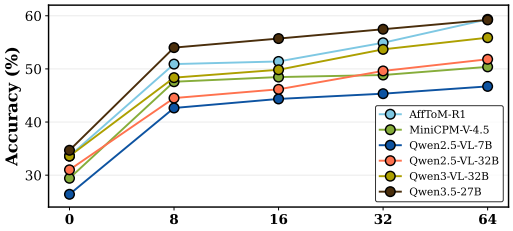


Figure 4: Accuracy across frame counts.

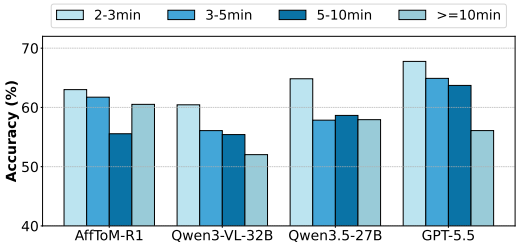


Figure 5: Accuracy across video durations.

to which the generated response remains faithful to the ground-truth context. Agreement between GPT-5-mini and human scores is assessed in Appendix C.3.

5.2 Main Results

We evaluate a diverse and representative set of MLLMs on the proposed benchmark. The overall performance metrics are summarized in Table 1, from which we derive several key and informative insights into current model capabilities. First, we can observe that MPAR is consistently the most challenging MCQ category. Even GPT-5.5, the strongest evaluated model, achieves only 52.05% accuracy on MPAR, substantially lower than its performance on the other question types. This gap highlights the difficulty of recursively tracking and integrating the affective states of multiple participants, exposing the limitations of current MLLMs in higher-order affective ToM reasoning. Second, larger models generally achieve stronger performance within the same family. This trend is observed in both MCQ and open-ended evaluations, suggesting that increased capacity is associated with better multimodal and socio-cognitive reasoning. The Qwen3.5 series is particularly competitive, with Qwen3.5-27B achieving the highest average scores among open-source MLLMs. Third, large-scale MLLMs yield further improvements across both evaluation formats. GPT-5.5 achieves the best overall performance on both MCQs and open-ended questions, demonstrating stronger capabilities in answer selection and free-form affective reasoning.

5.3 Effect of Frame Density and Video Duration

Impact of Frame Density. We evaluate the impact of frame density by varying the number of uniformly sampled video frames from 0 to 64, where 0 frames corresponds to question-only evaluation. The results are shown in Figure 4. First, transitioning from 0 to 8 frames yields a sharp performance increase across all models, showing that even sparse visual cues are essential for reliably inferring narrative trajectories. This upward trend generally continues up to 64 frames, further confirming the importance and effectiveness of denser spatiotemporal information. Second, model capacity strictly dictates the ability to process these dense inputs. While larger models consistently improve with more frames, smaller models plateau and exhibit slight fluctuations after 16 frames. This saturation indicates a representational bottleneck in smaller networks, whereas larger models possess the capacity required to effectively integrate and reason over extended temporal sequences.

Impact of Video Duration. We also investigate the effect of video duration, with results shown in Figure 5. Longer videos are generally more challenging, especially for GPT-5.5 and Qwen3-VL-32B, whose accuracy declines as the duration bucket increases. However, AffToM-R1 and Qwen3.5-27B exhibit non-monotonic changes across duration buckets, with partial recovery in the longest-video group. Among the evaluated models, GPT-5.5 achieves the best overall performance.

5.4 Analysis

Performance across Benchmarks. We evaluate the Qwen3.5-4B base model, its SFT variant, and AffToM-R1 on AffToM, Social-IQ 2.0 (Wilf et al., 2023), Hi-ToM (Wu et al., 2023), and the DX and Hard subsets of R3-Bench (Niu et al., 2026). These evaluations cover video-based social reasoning and text-based higher-order belief reasoning. Table 2 reports higher accuracy for both trained variants across all five evaluation sets, providing initial evidence that the benefits extend beyond AffToM. These gains suggest that learned reasoning transfers beyond the training distribution.

Table 2: Accuracy (%) across five benchmarks. Green values indicate gains over the base model in percentage points.

Dataset	Qwen3.5-4B	+SFT	AffToM-R1
AffToM	49.29	56.20 (+6.91)	**59.33** (+10.04)
Social-IQ 2.0	65.13	67.37 (+2.24)	**69.99** (+4.86)
Hi-ToM	43.75	49.33 (+5.58)	**50.25** (+6.50)
R3Bench-DX	70.99	75.55 (+4.56)	**75.74** (+4.75)
R3Bench-Hard	32.98	37.94 (+4.96)	**39.72** (+6.74)

Table 3: Ablation of PAFRO reward components on AffToM.

Reward Components			Accuracy (%)		
R_{persp}	R_{faith}	R_{cons}	IAR	MPAR	Overall
–	–	–	64.88	47.35	57.61
✓	–	–	64.92	48.08	57.94
✓	✓	–	65.04	49.80	58.73
✓	–	✓	65.32	49.01	58.56
✓	✓	✓	**65.51**	**50.60**	**59.33**

Ablation Study. We examine the contribution of these three reasoning rewards in PAFRO by enabling R_{persp}, R_{faith}, and R_{cons} while keeping the correctness and format rewards fixed. As shown in Table 3, R_{persp} provides the structural signal for reasoning from the required character viewpoint. Its effect is most visible on MPAR, where perspective control separates a character's mental state from the video's full information. Adding either R_{faith} or R_{cons} further improves the two-reward configuration, with faithfulness more closely supporting recursive mental-state tracking and consistency providing a stronger signal for direct affective inference. Faithfulness grounds mental-state claims in the reference graph, while consistency links evidence, reasoning, and the answer. The full reward set achieves the best results across IAR, MPAR, and Overall, reaching 65.51%, 50.60%, and 59.33%, respectively. These results support jointly enforcing perspective control, mental-state grounding, and evidence-answer consistency in PAFRO.

Effect of Information Revelation. Information asymmetry plays a critical role in shaping emotional dynamics. We conduct an analysis comparing model performance on pre-event and post-event questions, with results shown in Figure 6. Pre-event questions require models to simulate a restricted cognitive state and reason under incomplete information, where characters hold divergent beliefs or unresolved misunderstandings. In contrast, post-event questions evaluate affective states after the interaction unfolds and the epistemic uncertainty is largely resolved. We observe that models consistently achieve better performance on post-event questions. The higher proportion of Pre-N/Post-Y than Pre-Y/Post-N cases suggests that models are more likely to reason from an omniscient perspective, using later revealed information rather than preserving the character's pre-event knowledge state. These findings highlight a limitation of current models in handling information asymmetry, which is essential for robust affective ToM understanding. Additional experiments, error analyses, and representative case studies are provided in Appendix E.

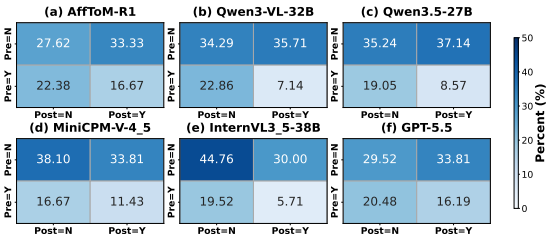


Figure 6: Joint correctness on paired MPAR questions. Rows show pre-event results and columns show post-event results. Y denotes a correct answer, and N an incorrect one.

6 Conclusion

In this paper, we propose AffToM, a pioneering multimodal benchmark specifically designed to evaluate affective ToM capabilities in dynamic multi-party social interactions. Moving beyond traditional epistemic reasoning and isolated emotion recognition, our dataset challenges models to decipher latent affective states within dynamic, multi-party social interactions. We construct the benchmark through an event-driven multi-agent generation pipeline grounded in dynamic character interaction graphs, followed by dual-criteria evolution and human validation to ensure data quality. To further investigate whether MPAR can be improved through targeted post-training, we introduce AffToM-R1, which combines SFT with perspective-aware rubric-guided optimization. Results across diverse MLLM families show that MPAR remains challenging, while AffToM-R1 delivers a substantial 13.38-percentage-point MPAR gain over its Qwen3.5-4B base model and consistent gains on external ToM and social-reasoning benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and to construct synthetic question-answer data for the benchmark and model training. For writing, generative AI tools were used to improve grammar, sentence structure, and readability. They were not used to formulate the research questions, design the benchmark, develop the methodology, or determine the experimental conclusions. For data construction, generative AI models were used within the multi-agent pipeline to generate candidate question–answer pairs, perspective labels, and reasoning trajectories from videos and structured character interaction graphs. These generated candidates were subsequently filtered through automated checks, iterative refinement, and human validation. The released benchmark annotations were reviewed and finalized by humans, rather than accepted solely on the basis of model-generated outputs. We reviewed all AI-assisted material and take responsibility for the final content of this work, including its text, claims, datasets, annotations, and other artifacts.

ETHICS STATEMENT

This work uses publicly available social videos and does not collect new personal data from human subjects. We use these videos in accordance with the terms of their original platforms and applicable copyright requirements. Our benchmark is intended for research on multimodal affective Theory of Mind and should not be treated as a reliable assessment of real individuals' emotions, beliefs, or intentions. The benchmark and its models are not intended for clinical diagnosis, employment, education, insurance, law enforcement, or other high-stakes decision-making. We do not attempt to identify or profile individuals in the source videos, and released materials will follow a research-only policy that prohibits re-identification, redistribution of source content, and inappropriate deployment.

REPRODUCIBILITY STATEMENT

The data generation pipeline is described in Section 3, and the training framework and reward design are detailed in Section 4. Appendix D reports the implementation settings, training configurations, and evaluation details. The prompts used for data generation, trajectory construction, refinement, and evaluation are provided in Appendix G. Together, these descriptions specify the data construction procedure, model training process, and evaluation protocol required to reproduce the reported experiments.

REFERENCES

Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, and Akshay Nambi. PromptWizard: Optimizing prompts via task-aware, feedback-driven self-evolution. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19974–20003, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1025. URL https://aclanthology.org/2025.findings-acl.1025/.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding, 2024. URL https://arxiv.org/abs/2411.04998.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding, 2026. URL https://arxiv.org/abs/2603.00512.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024.

Jean Decety and Philip L Jackson. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100, 2004.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

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 Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. 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. 5872–5877, 2019.

Ao Li, Longwei Xu, Chen Ling, Jinghui Zhang, and Pengwei Wang. Emoverse: Enhancing multimodal large language models for affective computing via multitask learning. *Neurocomputing*, 650:130810, 2025a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11929–11940, 2023. doi: 10.1109/ICCV51070.2023.01099.

Zhuo Li, Yuhao Du, Xiaoqi Jiao, Steven Y Guo, Yuege Feng, Xiang Wan, Anningzhe Gao, and Jinpeng Hu. Add-one-in: Incremental sample selection for large language models via a choice-based greedy paradigm. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5321–5340, 2025b.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9610–9614, 2023.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *International Conference on Machine Learning*, pp. 36993–37014. PMLR, 2025.

Angeline S Lillard. Pretend play skills and the child's theory of mind. *Child development*, 64(2): 348–371, 1993.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, 2025. doi: 10.1109/CVPR52734.2025.00315.

Meng Luo, Bobo Li, Shanqing Xu, Shize Zhang, Qiuchan Chen, Menglu Han, Wenhao Chen, Yanxiang Huang, Hao Fei, Mong-Li Lee, and Wynne Hsu. Unveiling the cognitive compass: Theory-of-mind-guided multimodal emotion reasoning, 2026. URL https://arxiv.org/abs/2602.00971.

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.

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 Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

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

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. URL https://arxiv.org/abs/2305.13786.

Li Qu, Pinxiu Shen, Yu Yan Chee, and Luxi Chen. Teachers' theory-of-mind coaching and children's executive function predict the training effect of sociodramatic play on children's theory of mind. *Social Development*, 24(4):716–733, 2015.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

ByteDance Seed. Seed1.8 model card: Towards generalized real-world agency, 2026a. URL https://arxiv.org/abs/2603.20633.

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity, 2026b. URL https://arxiv.org/abs/2607.00248.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind, 2025. URL https://arxiv.org/abs/2408.12574.

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

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22591–22611, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1230. URL https://aclanthology.org/2025.findings-emnlp.1230/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

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.

Yufan Wu, Yinghui He, 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. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717/.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.466. URL https://aclanthology.org/2024.acl-long.466/.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A Use of LLMs

We use LLMs for dataset construction, reasoning-trajectory generation, and rubric-based reward scoring. We evaluate pretrained models, train AffToM-R1 from a Qwen3.5-4B backbone through SFT followed by PAFRO, and use an LLM to score open-ended responses. LLMs assist with manuscript editing to improve grammar, sentence structure, and readability. They are not used to formulate research ideas or design the methodology. The authors take responsibility for the experimental conclusions and final manuscript.

B Dataset Details

B.1 Dataset Statistics

Video Sources and Duration. AffToM contains 258 videos collected from public video-sharing platforms, including 186 videos from Douyin/TikTok and 72 videos from YouTube. The dataset has a total duration of approximately 26.4 hours, with an average video length of 6.1 minutes. The collected videos primarily consist of social narratives, including short dramas and movie/TV drama clips, which naturally contain rich interpersonal interactions and evolving affective dynamics. The dataset includes videos in both Chinese and English, reflecting a range of social interaction contexts.

Scenario Diversity. As illustrated in Figure 7, AffToM covers a wide range of social scenarios rather than concentrating on a single interaction pattern. The collected videos span diverse interpersonal contexts, including family relationships, status/resource conflicts, hierarchical and political interactions, romantic relationships, peer friendships, and everyday social interactions. Such diversity provides rich social dynamics for evaluating affective Theory of Mind under realistic multi-party interactions.

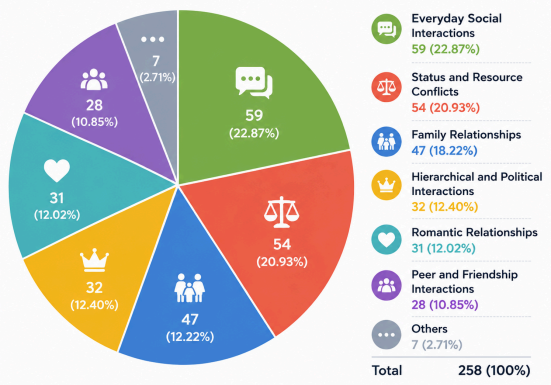


Figure 7: Scenario distribution of AffToM.

Filtering Statistics. During dataset construction, the multi-agent pipeline initially produced 4,174 question-answer pairs from 308 filtered videos. After automatic filtering for text-only shortcuts and option-level biases, 2,340 candidates were retained for human validation. At this stage, all QA candidates associated with 50 videos were filtered out. Consequently, these videos were excluded, reducing the dataset from 308 to 258 videos. Human annotators subsequently spent approximately 150 hours reviewing and validating the retained candidates.

Question Taxonomy and Distribution. After filtering and human validation, AffToM contains 1,922 question-answer pairs, including 1,822 multiple-choice questions and 100 open-ended questions. The questions are organized into two broad categories: Individual Affective Reasoning (IAR) and Multi-party Affective Reasoning (MPAR). IAR questions are presented in a multiple-choice format and cover six task paradigms: Preference Prediction, Integrated Mental-State Reasoning, Strategic Perspective-Taking, Hidden Intention, Personality Reasoning, and Counterfactual Reasoning. MPAR includes both multiple-choice and open-ended questions that evaluate recursive perspec-

tive tracking and belief-conditioned affective inference. Figure 8 presents the overall distribution of question types, together with their distribution across different video-duration ranges.

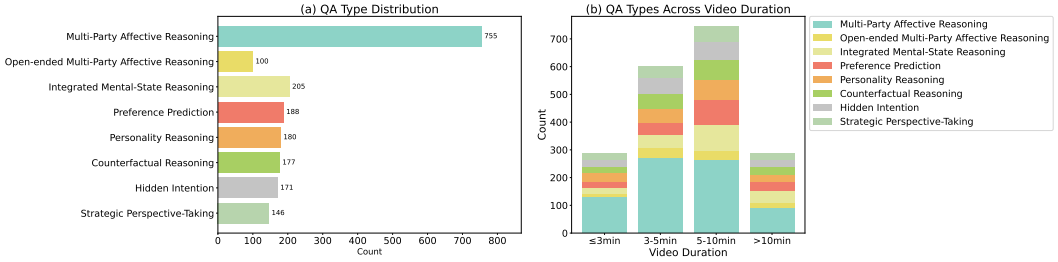


Figure 8: Distribution of QA types in the dataset and the distribution of QA types across different video-duration ranges.

B.2 Training Data Sources

The auxiliary training pool comprises 2,291 videos: 735 self-collected videos and 1,556 CondensedMovies clips from the R^3-FDT video collection (Niu et al., 2026). Applying our multi-agent data construction pipeline to these videos yields 19,759 QA samples, including 6,490 from the self-collected videos and 13,269 from CondensedMovies. We select subsets of this pool for SFT and PAFRO, with training settings detailed in Appendix D.1.

B.3 Comparison with Existing Benchmarks

Table 4 provides a detailed comparison between AffToM and representative social video and Theory-of-Mind benchmarks. We compare these datasets along three aspects: real-world video grounding, mental-state coverage, and advanced social reasoning capabilities. In this table, *Temporal Belief Update* refers to explicit evaluation of belief changes after new information is revealed, and *Recursive Perspective* refers to QA designs involving second-order or higher perspective recursion. Existing social video benchmarks, such as Video-MME, Social-IQ, MOMENTS, and R^3-Bench, provide valuable resources for evaluating general video understanding or social reasoning. However, they are not primarily designed to evaluate temporal belief updates or recursive perspective reasoning. MMToM-QA is more closely related to multimodal ToM evaluation, but is constructed in controlled environments and covers a more limited range of affective mental states. AffToM differs from these benchmarks by combining real-world social videos, multiple mental state dimensions, explicit update of temporal beliefs, and recursive perspective reasoning. In addition, AffToM includes multiple-choice and open-ended questions, which support evaluation in both discriminative and generative settings.

Table 4: Comparison with representative social video and Theory-of-Mind benchmarks. Temporal Belief Update denotes explicit belief changes after new information, while Recursive Perspective denotes second-order or higher perspective-recursive QA.

Benchmark	Real-World	Mental State				Temporal Belief Update	Recursive Perspective	Dataset Format
		Emotion	Belief	Intention	Desire			
Video-MME (Fu et al., 2025)	✓	✗	✗	✗	✗	✗	✗	MCQ
HI-TOM (Wu et al., 2023)	✗	✗	✓	✗	✗	✗	✓	MCQ
Social-IQ (Zadeh et al., 2019)	✓	✓	✓	✓	✓	✗	✗	MCQ
MMToM-QA (Jin et al., 2024)	✗	✗	✓	✓	✗	✗	✗	MCQ
MOMENTS (Villa-Cueva et al., 2025)	✓	✓	✓	✓	✓	✗	✗	MCQ
R^3-Bench (Niu et al., 2026)	✓	✓	✓	✓	✓	✗	✗	MCQ
AffToM	✓	✓	✓	✓	✓	✓	✓	MCQ + Open-ended

C Human Validation and Evaluation

C.1 Dataset Validation

Four students with backgrounds in computer science manually review the generated questions. Before large-scale annotation, they independently score a preliminary subset on four dimensions: an-

Table 5: Inter-annotator agreement measured by ICC(3,k) across evaluation dimensions. ANSR, VDEP, DQLT, and QVAL denote answer reasonableness, video dependency, distractor quality, and question validity, respectively.

Metric	ANSR	VDEP	DQLT	QVAL	AVG.
ICC (3,k)	0.811	0.823	0.852	0.810	**0.824**

swer reasonableness, video dependency, distractor quality, and question validity. Each dimension uses a 5-point Likert scale, ranging from 1 (poor quality) to 5 (superior quality). Substantial disagreements are discussed under the guidance of domain experts to establish a consistent evaluation standard. After calibration, all annotators independently score 100 randomly sampled instances to assess inter-annotator agreement. Table 5 reports ICC(3,k) scores above 0.80 across all four dimensions, indicating strong agreement. The annotators then verify the remaining question-answer pairs against the raw videos using the same four criteria, discarding any instance with an average score below 3.5.

C.2 Human Baseline Evaluation

To evaluate human performance on AffToM, we recruited five graduate students with backgrounds in computer science and multimodal learning as human testers. Although affective states may admit semantically related interpretations, each evaluated multiple-choice question was manually validated to contain a single best-supported answer grounded in the video context. Before the formal evaluation, all testers independently answered the same 30 randomly sampled questions without time constraints. Across this shared calibration subset, no pair of testers differed on more than three questions, indicating broadly consistent task interpretation. For the formal evaluation, we randomly sampled 250 multiple-choice questions and partitioned them evenly into five non-overlapping subsets, with each tester answering one subset. Thus, each question received one human response. Aggregating the responses across all subsets yielded 222 correct answers out of 250 questions, corresponding to an overall accuracy of 88.80%.

C.3 LLM-as-a-Judge Validation

To validate GPT-5-mini as the automatic evaluator for open-ended responses, a human annotator independently scores 100 responses generated by Qwen3.5-27B and another 100 responses generated by Qwen3-VL-32B using the same EC and FC rubrics. For Qwen3.5-27B, the Spearman correlations between human and GPT-5-mini scores are 0.766 for EC and 0.727 for FC. For Qwen3-VL-32B, they are 0.790 for EC and 0.769 for FC. These correlations support the use of GPT-5-mini for automatic scoring under the EC and FC rubrics.

D Experimental Implementation Details

D.1 Training Settings

We implement AffToM-R1 in ms-swift[1] using Qwen3.5-4B as the backbone. Both stages use AdamW, a constant learning rate schedule without warmup, bfloat16 precision, and a random seed of 42. Training inputs contain 16 video frames and the available subtitle text. Question difficulty is estimated from the pretrained backbone's accuracy over eight rollouts per question.

SFT optimization. Using the retained SFT subset, we train for two epochs with a learning rate of 2×10^{-5} and a batch size of 16.

PAFRO. The PAFRO policy and its fixed reference are initialized from the SFT-trained model. We train on 5k QA pairs for one epoch with a learning rate of 2×10^{-6} and a KL coefficient of 0.04. We use a batch size of 32 and generate four rollouts per question at temperature 1.0. Qwen3.5-27B serves as the reward model for the faithfulness and consistency rubrics. The reward weights $(\lambda_{\text{acc}}, \lambda_{\text{fmt}}, \lambda_{\text{persp}}, \lambda_{\text{faith}}, \lambda_{\text{cons}})$ are set to $(1.0, 0.1, 0.1, 0.4, 0.2)$.

[1] https://github.com/modelscope/ms-swift

D.2 SFT TRAJECTORY CONSTRUCTION

We construct each SFT trajectory with two successive calls to Qwen3.5-27B. In the first call, a perspective-classification prompt provides the video, question, and answer options, and asks the model to determine whether solving the question requires focused reasoning about the target character's information state and mental state. Questions that require character-specific beliefs or mental-state attribution are assigned the *character-centered* label, whereas questions that can be answered from the broader event context are assigned the *context-centered* label. In the second call, we provide the same video-question input together with the predicted perspective label and ask the model to generate a reasoning trajectory in the corresponding format. For *character-centered* questions, the trajectory follows *Clues* → *Mental State* → *Inference*. For *context-centered* questions, it follows *Clues* → *Inference*. The second prompt also constrains the response to include a `<perspective></perspective>` field, a reasoning block enclosed by `<think></think>`, and a final answer enclosed by `<answer></answer>`. All MPAR questions use the *character-centered* structure, while the remaining questions receive the structure selected in the first step. We retain trajectories whose final answers match the reference answers, and split the resulting approximately 20k-sample pool into SFT and PAFRO subsets according to the difficulty estimated from the pretrained backbone.

D.3 PAFRO OPTIMIZATION WITH GRPO

For each video-question pair $(v, q) \in \mathcal{D}_{\mathrm{RL}}$, the rollout policy π_{old} samples a group of K responses $\{y_i\}_{i=1}^{K}$. Each response receives the weighted reward r_i in Equation 9, which combines rule-based checks with the two rubric scores. We compare responses within the same group rather than fitting a separate value model. Writing $\bar{r} = K^{-1}\sum_{j=1}^{K} r_j$, the response-level advantage is

$$\hat{A}_i = \frac{r_i - \bar{r}}{\mathrm{Std}(r_1, \ldots, r_K) + \delta}, \tag{10}$$

where $\delta > 0$ stabilizes groups with little reward variation. Thus, a response is favored when its reward exceeds that of other responses to the same input.

The trainable policy π_θ and fixed reference policy π_{ref} are both initialized from the SFT-trained model, while π_{old} denotes the policy used to generate the current rollout group. For token $y_{i,t}$, let $\rho_{i,t} = \pi_\theta(y_{i,t} \mid v, q, y_{i,<t}) / \pi_{\mathrm{old}}(y_{i,t} \mid v, q, y_{i,<t})$ be the probability ratio between the updated and rollout policies. The clipped token objective is

$$\ell_{i,t}(\theta) = \min\Big(\rho_{i,t}\hat{A}_i,\ \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big), \tag{11}$$

where ϵ limits the size of policy changes. We maximize

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\left(\ell_{i,t}(\theta) - \beta d_{i,t}\right)\right], \tag{12}$$

where $d_{i,t}$ is the sampled token-level KL penalty relative to π_{ref}, and β sets its weight. The expectation covers training inputs and their sampled response groups. Responses receive equal weight, and tokens are averaged within each response. The same response-level advantage applies to all tokens of a rollout, linking the reward for its answer and reasoning to the policy update. Clipping and the fixed-reference penalty constrain the update while the group comparison favors higher-reward trajectories.

D.4 REWARD DEFINITIONS

This section details the five reward components combined in Equation 9.

Correctness reward R_{acc}. We compare the extracted answer $a(y)$ with the reference a^*:

$$R_{\mathrm{acc}}(y) = \mathbb{I}(a(y) = a^*), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function defined above. Invalid answer extractions receive zero reward.

Format reward R_{fmt}. This reward preserves the response structure learned during SFT. Let $\mathcal{F}$ denote responses that match the prescribed structure over the entire output. The canonical format is

<think>
<perspective>*perspective label*</perspective>
reasoning
</think>
<answer>*answer letter*</answer>

The <think> block must begin with the perspective field followed by reasoning, and the subsequent <answer> block must contain a single answer letter. Whitespace variations are permitted, but additional text outside these blocks is not allowed. This reward checks structural compliance. Perspective alignment and answer correctness are evaluated separately. We define

$$R_{\text{fmt}}(y) = \mathbb{I}(y \in \mathcal{F})\,. \tag{14}$$

Perspective reward R_{persp}. This reward compares the generated perspective label with the annotated route and target character. We extract the predicted label from the <perspective> field and normalize superficial differences in case and formatting. *Context-centered* questions require a *context-centered* prediction, while *character-centered* questions require a character-specific label sharing at least one character-descriptive token with the normalized reference. This lexical matching accommodates variation in character names and descriptions without requiring exact string equality. The reward is

$$R_{\text{persp}}(y) = \begin{cases} 1, & \text{the extracted perspective matches the annotated route,} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Missing or ambiguous perspective extractions receive zero reward.

Faithfulness reward R_{faith}. The reward model assesses the accuracy and relevance of mental-state claims by comparing them with the dynamic character interaction graph G provided as reference, using the sample's sub-event or whole-video scope. This reward applies to both routes, assessing *context-centered* reasoning without imposing a single-character viewpoint. For *character-centered* responses, it evaluates these claims from the viewpoint named in the generated label. We normalize the reward model's assessment under the faithfulness rubric to obtain

$$R_{\text{faith}}(y) = \text{Norm}(\mathcal{M}_{\text{faith}}(q_{\text{text}}, p(y), \tau_{\text{r}}(y), \mathcal{G}))\,, \tag{16}$$

where $\text{Norm}(z) = z/2$ maps the raw rubric score $z \in \{0, 1, 2\}$ to $[0, 1]$.

Consistency reward R_{cons}. The multimodal reward model evaluates whether the reasoning in y supports its final answer using evidence from video v and question q. It receives sampled frames and associated subtitles as scoring evidence, without the reference answer. The reward is

$$R_{\text{cons}}(y) = \text{Norm}(\mathcal{M}_{\text{cons}}(v, q, y))\,, \tag{17}$$

using the same normalization, with q including the answer options and y containing both the reasoning and the final answer. The rubric favors clear evidential support and penalizes contradictions or unsupported conclusions. The functions $\mathcal{M}_{\text{faith}}$ and $\mathcal{M}_{\text{cons}}$ denote separate calls to the reward model, each using its own rubric prompt to produce a response-level score.

D.5 Evaluation Settings

Evaluated Models. We evaluate a broad range of open-source MLLMs, including the Qwen2.5-VL series (Bai et al., 2025b), the Qwen3-VL series (Bai et al., 2025a), the Qwen3.5 series (Team, 2026), the Gemma-3 series, the InternVL3.5 series (Wang et al., 2025), MiniCPM-V-2.6 (Yao et al., 2024), MiniCPM-V-4.5 (Yu et al., 2025), models from the Ministral-3 series, LLaVA-OneVision (Li et al., 2024), and LLaVA-OneVision-1.5 (An et al., 2025). We also evaluate closed-source models, including GPT-5.5 and Seed-2.0-lite (Seed, 2026b).

Inference Settings. We set the decoding temperature to 0 during evaluation. For video inputs, we use 64 uniformly sampled frames per question unless otherwise specified. Open-source MLLMs are deployed and evaluated on NVIDIA RTX 5090 GPUs, while closed-source models are accessed through their vendors' commercial APIs.

Table 6: Effect of subtitle assistance on AffToM.

Models	Modality	MCQ			Open-ended		
		IAR	MPAR	Avg.	EC	FC	Avg.
Qwen3.5-4B	V	57.83	37.22	49.29	1.77	1.50	1.64
	V+T	**63.07**	**40.40**	**53.68**	**1.78**	**1.57**	**1.68**
AffToM-R1	V	65.51	50.60	59.33	**2.17**	1.89	2.03
	V+T	**66.35**	**52.72**	**60.70**	**2.17**	**1.96**	**2.07**
Qwen3.5-27B	V	69.26	45.03	59.22	2.45	2.23	2.34
	V+T	**73.85**	**51.79**	**64.71**	**2.69**	**2.66**	**2.68**
GPT-5.5	V	71.70	52.05	63.56	3.06	2.99	3.03
	V+T	**76.10**	**56.95**	**68.17**	**3.51**	**3.46**	**3.49**

Table 7: Effect of different frame sampling strategies on MCQ performance. We report both overall accuracy and MPAR accuracy (%).

Models	Uniform	BOLT (Liu et al., 2025)	WFS-SB (Chen et al., 2026)
Qwen3-VL-32B	55.87 / 40.79	57.85 / 44.90	56.31 / 44.37
Qwen3.5-27B	59.22 / 45.03	61.80 / 49.67	60.76 / 48.21

E Additional Experiments and Analysis

E.1 Effect of Subtitles Assistance on AffToM

We augment video inputs with subtitles generated by Whisper *large-v2* (Radford et al., 2022). Table 6 compares video-only inputs (V) with video-plus-subtitle inputs (V+T) across the evaluated models. Subtitle assistance improves MCQ accuracy for every reported model, with the largest gains generally appearing on MPAR. Its effects on open-ended EC and FC scores are mixed across models, with improvements observed for some larger baselines. For the Qwen3.5-4B base model and AffToM-R1, adding subtitles improves MCQ accuracy by 4.39 and 1.37 percentage points, respectively. These results indicate that dialogue information provides complementary evidence for interpreting characters' beliefs and emotions.

E.2 Effect of Frame Sampling Strategy

We further investigate whether adaptive frame selection can improve model performance under a fixed frame budget. Specifically, we replace uniform sampling with BOLT (Liu et al., 2025) and WFS-SB (Chen et al., 2026), while keeping all other inference settings unchanged. As shown in Table 7, both strategies consistently outperform uniform sampling for Qwen3-VL-32B and Qwen3.5-27B. BOLT achieves the strongest results, improving overall MCQ accuracy by 1.98 and 2.58 percentage points and MPAR accuracy by 4.11 and 4.64 points, respectively. WFS-SB also yields consistent, though relatively smaller, improvements across both metrics. Notably, the gains on MPAR are consistently larger than those on the overall MCQ set, suggesting that adaptive sampling is particularly beneficial when the required evidence is concentrated in a small number of socially decisive moments. By reducing visually redundant frames and prioritizing more informative temporal evidence, these strategies facilitate the tracking of belief updates and inter-character dependencies.

E.3 Error Analysis

To further characterize model failures, we analyze the generated reasoning outputs of Qwen3.5-27B and Qwen3-VL-32B on their respective incorrectly answered MPAR MCQs and classify the resulting failures into four categories: *Temporal Information Leakage*, *Recursive Perspective Conflation*, *Unwarranted Affective Deduction*, and *Others*, as summarized in Figure 9. For **Pre-Event** questions, *Temporal Information Leakage* is the most prominent error, indicat-

ing that models tend to rely on information revealed at later event stages when reasoning under asymmetric information. For **Post-Event** questions, *Unwarranted Affective Deduction* becomes the dominant error type. This shift demonstrates that while models can observe the physical event, they struggle to accurately map these visual observations to updated psychological states, frequently hallucinating unsupported affective dependencies or over-interpreting sparse cues. For **Non-pair** questions, *Recursive Perspective Conflation* emerges as the most severe bottleneck. Absent explicit temporal anchoring, models persistently conflate objective multimodal facts with the target character's subjective, nested beliefs. Notably, despite differences in overall model capacity, Qwen3.5-27B and Qwen3-VL-32B exhibit highly consistent error distributions across all three temporal conditions. Overall, these findings indicate that higher-order affective reasoning is mainly limited by temporal grounding under asymmetric information, faithful recursive perspective tracking, and psychologically plausible inference.

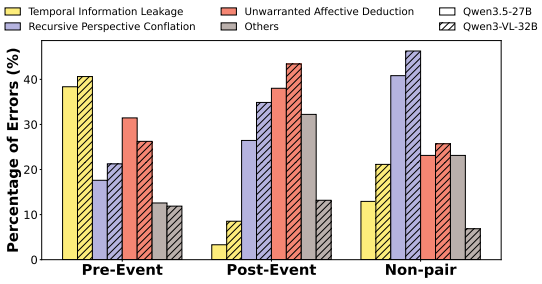


Figure 9: MPAR MCQ error distribution (%). Non-Pair denotes questions outside the pre/post-event pairs.

E.4 Case Study

Figure 10 presents representative failures that illustrate the error patterns above and the degradation on open-ended questions. In Figure 10 (a), the model's reasoning trace reveals a clear temporal information leakage error. Although the question is anchored before any spray is used, the model justifies its answer by referring to the woman's later reaction, namely that she sprays the man and calls him a bad person. This evidence is unavailable at the queried moment: before the spray, the man has not yet inferred that the woman mistakes him for a robber or feels threatened by him. By using this post-event reaction to infer the man's pre-event belief, the model answers from an omniscient timeline rather than from the man's temporally constrained perspective.

In Figure 10 (b), the model's reasoning trace reveals a recursive perspective conflation error. The question targets the boyfriend's sister's belief about the girlfriend's thoughts and feelings, but the model grounds its answer in the girlfriend's own latent dissatisfaction. It cites Chu Chu's remarks about needing a ceremonial wedding and her concerns about the house and betrothal gift, and then infers that the boyfriend's sister's tone and words show she perceives Chu Chu's disappointment. In this way, the model treats the girlfriend's internal disappointment as if it were already represented in the sister's belief. The predicted answer therefore conflates the girlfriend's true inner state with the sister's second-order perspective, rather than reasoning only from what the sister can observe and infer.

In Figure 10 (c), the model's reasoning trace exposes an unwarranted affective deduction rather than a temporal or perspective-tracking error. The model stays within the Emperor's current viewpoint, but overinterprets his dialogue and tone. From his concern and the utterance "You don't want to see me again?", it infers that he believes the concubine is avoiding him out of guilt, sorrow, self-punishment, or a desire to spare him pain. This goes beyond the available evidence: the scene only supports that she claims illness and that the Emperor responds with concern for her condition. The failure is therefore an unsupported evidence-to-emotion mapping, where sparse cues are turned into a strong but unjustified emotional-causal chain.

In Figure 10(d), the model initially preserves the outer belief chain but partially collapses the recursive attribution in its subsequent explanation. Specifically, the final statement directly attributes the girlfriend's emotional state to the woman in brown, rather than maintaining the woman's belief about the boyfriend's inference. The response therefore represents the recursive structure only incompletely.



Figure 10: Representative failure cases of MLLMs on AffToM, including three major error types in MPAR—Temporal Information Leakage, Recursive Perspective Conflation, and Unwarranted Affective Deduction—as well as a typical error in open-ended question answering involving missing recursive characters.

F MULTI-AGENT DATA CONSTRUCTION

F.1 SYSTEM CONFIGURATION

Our multi-agent system is implemented using the LangGraph framework [2] and consists of six agents: the Event Agent, State Agent, Question Generation Agent, Role Reflection Agent, Revision Agent, and Evaluation Agent. The Event Agent and State Agent extract structured events and character-level mental states, which are integrated into G for subsequent question generation and refinement. Based on preliminary experiments, we use Seed-1.8 as the backbone of the Event Agent and State Agent due to its stronger semantic extraction capability, while the remaining reasoning and refinement agents use Qwen3-VL-32B as their backbone.

All agents use *top_p* $= 0.95$. The Event Agent, State Agent, and Evaluation Agent use *temperature* $= 0$ for deterministic extraction and evaluation, while the remaining agents use *temperature* $= 0.7$ for diverse generation, reflection, and revision. During iterative refinement, we randomly sample $|\mathcal{Q}_{\text{human}}| = 5$ human-authored reference questions from the seed pool for pairwise comparison. We set the acceptance threshold to $\tau = 0.8$, the trade-off coefficient to $\lambda = 0.7$, and the maximum number of refinement rounds to $T_{\max} = 3$. For recursive perspective expansion, we set the maximum recursive depth to $d_{\max} = 3$, balancing reasoning complexity and annotation reliability.

F.2 MPAR QUESTION GENERATION EXAMPLE

F.2.1 EVENT EXTRACTION OUTPUT

Here we present an example of an event list extracted from a single video. It records the main characters, event timestamps, locations, event summaries, and involved individuals. The event list directly provides the event-level context for base-level question generation and, together with the corresponding individual-level information, is used to construct the dynamic character interaction graph G for recursive perspective expansion.

Event List Example:


```
"character_list": [{
  "character_id": "Wang Zha",
  "description": "A young man in a black jacket, trying to
      reconcile with his girlfriend Leng Meng and later proposing to
       her. His parents misinterpret his interactions with his
      friend."
},
{
  "character_id": "Leng Meng",
  "description": "Wang Zha's girlfriend, wearing a white coat,
      sitting upset on a park bench after a misunderstanding with
      Wang Zha."
},
{
  "character_id": "Leng Meng's male friend",
  "description": "A man in a denim jacket with glasses, protecting
      Leng Meng from Wang Zha initially, and later mistakenly
      targeted by Wang Zha's parents."
},
{
  "character_id": "Wang Zha's father",
  "description": "An older man in a gray jacket, who misinterprets
      Wang Zha's interactions with his male friend as a romantic
      relationship and becomes distraught."
},
{
```


[2] https://github.com/langchain-ai/langgraph

```
    "character_id": "Wang Zha's mother",
    "description": "An older woman in a beige coat, who tries to calm
        her husband and also misinterprets Wang Zha's interactions,
       later offering money to Leng Meng's friend."
  }],
  "events": [
  {
    "event_id": "event_1",
    "timestamp": "00:00 - 00:37",
    "location": "Outdoor park area",
    "description": "Wang Zha tries to approach the upset Leng Meng,
       but her male friend blocks him. Wang Zha argues with the
       friend, using casual terms like 'dear' which are overheard by
       Wang Zha's parents. The parents misinterpret the interaction
       as Wang Zha being in a romantic relationship with the male
       friend.",
    "interaction_pairs": [
      ["Wang Zha", "Leng Meng's male friend"],
      ["Leng Meng's male friend", "Wang Zha"],
      ["Leng Meng's male friend", "Leng Meng"],
      ["Wang Zha", "Leng Meng"],
      ["Leng Meng", "Wang Zha"],
      ["Wang Zha's father", "Wang Zha's mother"],
      ["Wang Zha's mother", "Wang Zha's father"],
      ["Wang Zha's father", "Wang Zha"],
      ["Wang Zha's mother", "Wang Zha"]
    ]
  },
  ...]
```


F.2.2 Graph Initialization and Recursive Question Generation

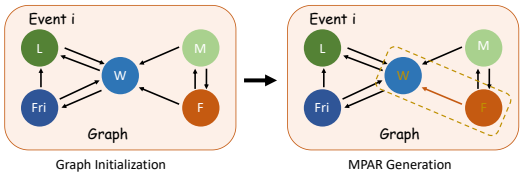


Figure 11: An example of a character interaction graph. The nodes represent "L": "Leng Meng", "Fri": "Leng Meng's male friend", "W": "Wang Zha", "M": "Wang Zha's mother", and "F": "Wang Zha's father". The left panel shows the initialized graph, while the right panel highlights the characters selected for recursive perspective construction during MPAR generation.

Based on the extracted events $\mathcal{E}$ and individual-level information $\mathcal{S}$, we construct the dynamic character interaction graph G, as illustrated in Figure 11. Each node represents an involved character, while each directed edge indicates the direction of an observed interaction. The left panel of Figure 11 shows the initialized graph, and the right panel illustrates how it supports Recursive Perspective Expansion for MPAR question generation.

We first establish a base-level affective state or belief for a target character grounded in the event context. In the running example, Wang Zha is selected as the target character. To expand the reasoning structure, we identify characters who directly interact with or observe the current character within the current event context encoded in G_i and treat them as candidate outer observers. When multiple candidates are available, the LLM jointly examines the graph G_i, the event-localized video segment, and the current recursive belief chain. The selection considers whether the event supports information asymmetry between the candidate and the current subject, whether the candidate's belief concerns the current subject, and whether the associated emotion, intention, or desire is relevant to

the target relation. The selected character is then introduced as the new outer observer, extending the recursive perspective chain.

In Figure 11, Wang Zha's father, represented by node F, is selected as the outer observer. This expands the reasoning structure from Wang Zha's base-level mental state to a second-order perspective. The graph G_i, the corresponding video segment, and the current belief chain are subsequently provided to the Question Generation Agent, which produces an MPAR question-answer pair together with plausible distractors.

F.3 Pipeline Robustness Audit

To further assess the robustness of the multi-agent construction pipeline, we conduct an additional audit of representative intermediate outputs. During dataset construction, intermediate supervision is incorporated into the pipeline: before large-scale generation, we inspect stage-level outputs and proceed only after the outputs become stable. To quantitatively examine whether upstream errors could systematically affect final QA quality, we further re-audit 20 randomly sampled generation traces across three stages that are most relevant to final QA validity: event extraction, role/mental-state extraction, and recursive QA construction.

We apply stage-specific criteria to inspect intermediate outputs at the level of individual checkable positions. Each occurrence that violates a criterion contributes one error count, and the error rate is computed as the total number of errors divided by the total number of corresponding positions inspected across the 20 generation traces. For event extraction, the inspected positions include character-interaction edges, event temporal boundaries, character identities, and required output format. Errors are counted for incorrectly constructed interaction edges, temporal intervals that substantially deviate from the event's reasonable boundaries, misidentified characters, or violations of the prescribed output format. For role/mental-state extraction, we inspect character profiles, event-specific mental-state entries for each present character, and required output format. Errors include inaccurate profile information, missing mental-state entries for characters involved in an event, mental-state analyses that are clearly inconsistent with the video evidence, and formatting violations. For recursive QA construction, we inspect every belief-extraction step along the recursive perspective chain, together with the required output format. Each incorrect belief extraction or formatting violation contributes one error count.

As shown in Table 8, most inspected intermediate positions are consistent with the video, although residual errors remain across all three audited stages. Role/mental-state extraction shows an error rate of 5.22%, reflecting occasional inaccurate mental-state analysis in sub-events. Recursive QA construction has an error rate of 7.41%, mainly caused by unsupported relational or affective links. Given the limited sample size and the position-level nature of this audit, these results should be interpreted as a diagnostic analysis rather than a dataset-wide reliability estimate. The observed errors, particularly unsupported relational or affective links, are further screened during final human validation, where annotators verify, revise, or discard QA candidates based on the original video evidence and event-localization information.

Table 8: Pipeline robustness audit on 20 randomly sampled generation traces. Error rates are computed over all inspected intermediate positions.

Stage	Primary Observed Error	Error Rate
Event Extraction	Incorrect interaction edges	3.67%
Role/Mental-State Extraction	Incorrect mental-state analysis in events	5.22%
Recursive QA Construction	Unsupported relational/affective links	7.41%

G PROMPT DETAILS

G.1 PROMPT FOR EVENT INFORMATION

Instruction:
You are an expert in video analysis and information extraction. Your goal is to extract important events from the video and identify the interaction pairs of characters in each event, in order to construct a directed graph of character interactions.

 # Task Overview:
Complete the task in two stages:
1. Identify the main characters.
2. Extract valid events and their interaction pairs.

Stage 1: Character Identification
Identify the main characters involved in the plot.
Character ID Requirements:
Assign each character a natural-language unique ID.
Prohibited Naming Patterns:
Do not:
1. Use underscores ('_'), suffix numbers (such as 1 or 2), or code-like variables.
2. Combine multiple descriptors into one ID.
3. Include personality traits or character judgments. **### Character Naming Strategy:**
Use the following priority order when assigning character IDs:
1. The character's proper name, if it is mentioned or known.
2. A single relationship identifier, such as husband, wife, or boss.
3. A single distinctive visual feature, such as clothing or hairstyle, only if name and relationship are not available.

Examples
- BAD: "friend_1" (number suffix), "Man_in_Blue" (underscore), "the kind Boss" (personality/trait), "the son, young boy" (multiple descriptors)
- GOOD: "Xiao Ming" (proper name), "the husband" / "the Boss" (single relationship), "the woman in the blue vest" (single visual feature)

Stage 2: Event Extraction
Extract a list of events that strictly satisfy the requirements below.
Core Constraint:
Each extracted event must occur within a single continuous scene.
Prerequisite — Information Asymmetry:
Before extracting any event, determine whether the scene contains an information gap between characters.
Only extract events where information asymmetry exists.
Information asymmetry occurs when different characters possess different knowledge about a situation, leading to divergent beliefs, understandings, or expectations.
Examples of Information Asymmetry:
1. Hidden actions: Alice and Bob both see apples in a basket. After Bob leaves, Alice takes the apples. Bob is unaware of this action and still believes the apples are in the basket.
2. Hidden facts or intentions: Alice's birthday is approaching. Alice firmly believes Bob has not prepared a gift for her, but Bob actually has one hidden away.
Interaction Pair Extraction:
For each valid event, provide 'interaction_pairs': specific pairs of characters who have explicit, observable interactions within that event.
An interaction must satisfy at least one of the following criteria:
1. Verbal communication: direct dialogue, arguing, or targeted vocal exchanges.
2. Physical or object interaction: physical contact, such as pushing or hugging, or handing/receiving objects.

3. Non-verbal communication: mutual eye contact, targeted facial expressions, or gestures.
4. Joint action or observation: collaborating on a shared task, or one character deliberately watching or eavesdropping on another.
Interaction Pair Formatting Rules:
1. Use the exact Character IDs defined in Stage 1.
2. The edges are directed.
3. If the interaction is one-sided observation, for example A secretly observes B, output it as '[A, B]'.
4. If the interaction is mutual, output both pairs: '[A, B]' and '[B, A]'.
5. Do not assume everyone interacts with everyone.
6. If A simply walks past B without meeting any of the four criteria above, do not include that pair.

Field Definitions:
"description": Use a brief plain-language description. Keep it simple and focused on the most salient observable content.
"timestamp": Use a start-end time span that covers the full event at the event level, in the format "MM:SS - MM:SS".
"location": Use an abstract scene-level location label (e.g., "kitchen", "park area", "office hallway"), rather than fine-grained coordinates or camera positions.

Output Format:
The final output must strictly follow this JSON format:

```
{
  "character_list": [
    {
      "character_id": "Alice",
      "description": "..."
    }
  ],
  "events": [
    {
      "event_id": "event_1",
      "timestamp": "MM:SS - MM:SS",
      "location": "...",
      "description": "...",
      "interaction_pairs": [
        ["Alice","the mother"],
        ["Alice", "the man in black coat"]
      ]
    }
  ]
}
```

G.2 Prompt for Character Mental-State Extraction

Instruction:
You are a psychology expert specializing in Mental State Representation (MSR) modeling. Your core mission is to bridge the gap between observable behavior and internal cognition. Based on objective event descriptions, you must deeply analyze the psychological states of characters within various event-driven conflicts.
Please conduct in-depth analysis using the following Theory of Mind (ToM) classification system: {TOM_KNOWLEDGE_BASE}
Construct causal chains to predict character behavior and verify the chains against video content: {TOM_CAUSAL_CHAIN}

Task Requirements: Multi-dimensional Analysis
1. For each event in the "Event List" below, locate the corresponding video segment, analyze the mental states of each involved character, and add the analysis to $event_i$.
2. Correct character ToM modeling must align with the video content. Perception is critical: incorrect perception invalidates subsequent belief, desire, intention, and emotion analysis. After establishing the initial perception, conduct reflective verification through detailed comparison with the video content before analyzing subsequent mental states.
3. If multiple similar ToM mental states appear within the same event, list them in chronological order.
4. Based on the extracted information, summarize each character's personality profile, including Occupation, Character Relationships, and Preferences to Objects.
- Record profile attributes only when they are explicitly stated in the video or consistently supported by multiple cues across the video. Otherwise, set the corresponding field to null.
- <Occupation> extracts the character's job if it exists. Otherwise, set it to null.
- <Character Relationships> extracts only relationships where the character has interacted with another character at least once in some event.
- <Preferences to Objects> extracts content related to the character's preferences, needs, or attitudes toward objects or concepts, e.g., likes money or dislikes apples.
Character List & Event List: {event_list_json}
The final output must strictly adhere to the following JSON format:

```
{
  "Character A": {
    "profile": "Basic information and feature preferences. If
       present, include <Occupation>, <Character Relationships>,
       and <Preferences to Objects>.",
    "event_1": {
      "timestamp": "00:00 - xx:xx",
      "location": "Location name, e.g., Kitchen",
      "appearance": "Character appearance, e.g., wearing a red
         shirt",
      "mental representation": {
        "Percept": "...",
        "Belief": "...",
        "Desire": "...",
        "Intention": "...",
        "Emotion": "..."
      }
    },
    ...
  },
  "Character B": {
    "profile": "...",
    "event_1": {
      "...": "..."
    }
  }
}
```

G.3 Prompt for Role Reflection and QA Evolution Advice

Instruction:
You are the "Character Simulator" and the "QA Evolution Advisor".
Your goal is to progressively evolve a QA pair into a high-quality mentalizing challenge.
Role Playing (Target Character: {character_name})
We provide you with the character interaction graph, which contains the interaction structure and character-level mental-state representations, to simulate the target character's perspective.

- **Independent Perspective Simulation**:
Answer the question independently in the first person from {character_name}'s perspective, using the interaction structure and character-level mental-state information encoded in the graph. Do not treat the wording of any answer option as evidence.
- **Consistency Check**:
After forming this independent judgment, compare it with the current correct answer. Determine whether the current answer is consistent with the target character's perspective and identify any discrepancy.
- **Output (`role_reflection`)**:
Summarize the independently inferred perspective and the result of the consistency check in `role_reflection`.
Advisor Role
Provide modification suggestions based on the independent role reflection, the current QA, and the critique from the professional Evaluator.
Input Data:
- Current QA:
{current_qa_json}
Current QA Type Definition
{qa_type_definition}
- Your Mental State & Interaction Information:
{graph}
- Potential Modification Critique:
{evaluator_critique}
Tasks:
- **Strategy Application**:
Rewrite distractors using the following strategies:
{distractor_strategies}
- **Handling Reflection and Critique**:
Use the independent role reflection to examine whether the current correct answer faithfully represents the target character's perspective. Incorporate useful evaluator feedback when revising the distractors, answer logic, or question formulation.
- **Output (Modification Suggestions)**:
Clearly describe how to improve the distractors, answer logic, or question formulation while preserving the original reasoning objective and video-grounded evidence.
The final output must strictly adhere to the following JSON format:

```
{
  "role_reflection": "An independent first-person perspective and
      its consistency with the current correct answer.",
  "modification_advice": "Detailed but concise suggestions for
      improving the QA pair without directly rewriting the complete
      QA."
}
```

G.4 Prompt for QA Revision Agent

Instruction:
You are the "QA Revision Agent".
Your goal is to refine the QA pair into a higher-quality mentalizing challenge based on simulator feedback and video evidence.
Revision Role
We provide the original QA pair, the current QA type definition, simulator feedback, and optimization history.
Your task is to revise the QA pair while preserving the original reasoning objective and improving its logical coherence, video dependency, and distractor quality.
Input Data:

- **Original QA:**
{qa_to_improve_json}
Current QA Type Definition
{qa_type_definition}
- **Simulator Feedback:**
- `role_reflection`: The target character's internal perspective regarding the current correct answer.
- `modification_advice`: Suggestions for improving distractors, answer logic, or question quality.
{simulator_feedback}
- **Optimization History:**
{optimization_history}
Tasks:
- **Video Grounding**:
Verify all revisions against the original video content and ensure that the revised QA remains grounded in observable evidence.
- **Feedback Integration**:
Incorporate the simulator feedback to improve the QA pair.
a. Revise the correct answer when necessary to better align with the target character's perspective.
b. Rewrite or replace distractors according to the provided modification advice, ensuring that they remain plausible but incorrect.
- **Quality Control**:
Ensure that all answer options have comparable length and maintain the original `question_type`.
- **Candidate Generation**:
Generate {num_candidate_versions} candidate versions with diverse but logically valid revisions.
The final output must strictly adhere to the following JSON format:

```
{
  "questions": [{
      "question": "...",
      "options": {
        "A": "Revised Option A...",
        "B": "Revised Option B...",
        "C": "Revised Option C...",
        "D": "Revised Option D..."
      },
      "correct_answer": "...",
      "question_type": "(Keep Original Type)",
      "timestamp": "..."
    },
    ...
  ]
}
```

G.5 Prompt for Multiple-Choice Questions

 # Instruction:
These are the frames of a video.

(Optional) The video's subtitles are provided below when subtitle assistance is enabled:

{Subtitles}

Select the best answer to the following multiple-choice question. Respond with the exact final sentence "`<answer>X</answer>`", where X is one of A, B, C or D.

 # Question: {Question}

 # Options:
A: {Option A}
B: {Option B}
C: {Option C}
D: {Option D}

 # Answer:

G.6 Prompt for Open-Ended questions

 # Instruction:
These are the frames of a video.

(Optional) The video's subtitles are provided below when subtitle assistance is enabled:
{Subtitles}
Please answer the following question.

 # Question: {Question}

 # Task:
Your answer must explicitly describe a three-level belief structure. Specifically, you should express Agent A's interpretation of Agent B's belief about Agent C's perspective on the event, as well as Agent A's inference of Agent B's belief about Agent C's emotional attitude toward the event.

G.7 Prompt for LLM-as-a-Judge in Open-Ended questions

 # Instruction
You are a helpful assistant in evaluating open-ended Q&As. Please act as an unbiased judge. The reference answer has been manually validated against the source video and serves as the factual anchor for evaluation. Evaluate the assistant's response based only on the question and the human-validated reference answer. Assess whether the response preserves the factual content, character roles, and recursive viewpoint structure represented in the reference answer. Do not introduce or reward additional interpretations solely because they appear plausible.

 # Notes
- Your evaluation should determine whether the assistant effectively follows the requirements of the question and correctly addresses it.
- You must assess the response across the two dimensions and score each dimension independently.
- Do not let the length of the response influence your judgment.
- Do not favor any specific assistant identity or position. Remain as objective as possible.
- Do not penalize wording differences when the response is semantically equivalent to the reference answer.

 # Criteria
Use scores to represent the quality of the assistant's response. Below are the detailed scoring rubrics.

Emotion Correctness (EC)
This dimension evaluates whether the response correctly identifies the emotional state of the target character, consistent with the reference answer. Factual alignment of the emotion is the primary criterion.
- **1 – Poor:** The predicted emotion is entirely opposite to the reference answer. Even if the nested perspective is grammatically fluent, an opposite emotion must receive the lowest

score.
- **2 – Fair:** The predicted emotion shows weak relevance and contains noticeable errors, such as identifying an incorrect or overly weak emotion.
- **3 – Moderate:** The general emotional direction is somewhat correct, but lacks clarity, specificity, or accurate recursive attribution.
- **4 – Good:** The emotional polarity is highly consistent with the reference answer, with only minor differences in wording or intensity.
- **5 – Excellent:** The predicted emotion fully matches the reference answer in polarity, intensity, and semantic precision, and is correctly attributed within the recursive belief structure.

Factual Consistency (FC)
This dimension evaluates the accuracy of the embedded viewpoint content and whether the response correctly represents the recursive belief structure (i.e., Agent A's perspective on Agent B's belief about Agent C's view).
- **1 – Poor:** The core viewpoint contradicts the reference answer or the recursive chain is missing or incorrect.
- **2 – Fair:** The viewpoint is largely inaccurate, focused on the wrong event, or severely misinterprets the characters' intentions.
- **3 – Moderate:** The recursive structure is identifiable, but the viewpoint only partially aligns with the reference and contains omissions or minor role confusion.
- **4 – Good:** The recursive chain is clear and correctly ordered. The viewpoint closely matches the reference answer, with only minor semantic differences.
- **5 – Excellent:** The recursive structure is explicit, correctly ordered, and unambiguous. The embedded viewpoint perfectly matches the reference answer in meaning, focus, and interpretation.

 # Test Question
QUESTION: {question_text}
ASSISTANT ANSWER: {pred_answer}
GROUND TRUTH: {ground_truth}

 # Output Format
You MUST output your evaluation strictly in the following JSON format:

```
{
    "Emotion_Correctness": <score from 1 to 5>,
    "Factual_Consistency": <score from 1 to 5>
}
```