ToM-Video: A Theory-of-Mind Video Benchmark with Spatio-Temporal Grounded Reasoning

Anonymous ACL submission

Abstract

Humans can infer others' beliefs, intentions, and emotions from observable behavior, an ability known as Theory of Mind (ToM). While current Multimodal Large Language Models (MLLMs) excel at visual understanding, can they identify mental-state cues in videos and reason about evolving mental states? To systematically evaluate spatio-temporal ToM in MLLMs, we introduce ToM-Video, a large-scale benchmark built from diverse real-world narrative videos. Through multi-stage filtering and annotation from massive data sources, ToM-Video provides a high-quality collection of ToM-centric scenes, comprising 1,487 videos and 41K visual question answering (VQA) pairs. We further propose INSPIRE, a post-training framework that, through a carefully designed three-stage curriculum, progressively teaches models to locate ToM-relevant cues and connect them to latent mental-state changes. Experiments show that current MLLMs remain weak at linking observable cues to evolving mental states, while INSPIRE shifts the model from surface description toward compact, evidence-backed inference, leading to higher accuracy and better generalization to external social-reasoning video datasets.

1 Introduction

The ability to infer the underlying mental states from observed behavior is known as *Theory of Mind* (ToM) (Premack and Woodruff, 1978; Beaudoin et al., 2020). Serving as a significant non-linguistic complement to prevailing language-dependent agent interaction paradigms, ToM presents a crucial step towards human-AI cooperation, enabling agents to better understand human perceptions, intentions, and expectations while interacting in shared environments.

Recent multimodal large language models (MLLMs) have achieved remarkable advances in general video understanding, bringing promising potential for tackling ToM reasoning (Zhang et al., 2025; Bai et al., 2025a). Nevertheless, ToM-centric tasks remain severely underdeveloped in current video model training and evaluation paradigms. Unlike object or action recognition, ToM depends on how latent mental states are formed and revised over time (Baker et al., 2017; Rabinowitz et al., 2018; Jara-Ettinger, 2019). Existing multimodal ToM benchmarks, as summarized in Table 1, have begun to expose this problem, but each addresses only part of the required evidence structure. They often compress video evidence into final answers or textual causal descriptions. As a consequence, two indispensable signals are erased at the source: (I) As mentioned above, ToM reasoning tasks are highly dependent on time, the dense *temporal* structure that only the raw video can provide; (II) the fine-grained separation between the *audio* and *visual* streams (Chakraborty et al., 2025), which in real social interactions are frequently incongruent (e.g., a smiling face paired with a sarcastic tone).

Bench.	*Setting*		*Scope*	*Grd.*		*Sup.*	#QA
	Video	Real	Type	T	Bx	G	
ToMBench (Chen et al., 2024b)	✗	✗	6	✗	✗	✗	2.9K
MMToM (Jin et al., 2024)	✓	✗	2	△	✗	✗	600
MuMA-ToM (Shi et al., 2025)	✓	✗	2	✗	✗	✗	900
Social-IQ2 (Wilf et al., 2023)	✓	✓	Imp.	✗	✗	✗	6K
IntentQA (Li et al., 2023)	✓	✓	1	✗	✗	✗	16.3K
VToM (Chen et al., 2025b)	✓	✓	4	△	✗	✗	6K
MoMentS (Villa-Cueva et al., 2025)	✓	✓	7	✗	✗	✗	2.3K
EgoToM (Li et al., 2025b)	✓	✓	2	△	✗	✗	1.0K
R^3-VQA (Niu et al., 2026)	✓	✓	4	✗	✗	△	5.1K
ToM-Video	✓	✓	**7**	✓	✓	✓	**41.2K**

Table 1: Comparison with representative ToM and social-reasoning benchmarks. **Scope**: **Type counts major ToM-relevant categories targeted by each benchmark** (See appendix A for details). Imp. = implicit social-state reasoning without an explicit ToM taxonomy. **Grd.**: T = timestamp, or event-time grounding; Bx = keyframe-level bounding-box grounding. **Sup.**: G = grounded reasoning-chain supervision. ✓, △, and ✗denote full, partial, and no explicit support.

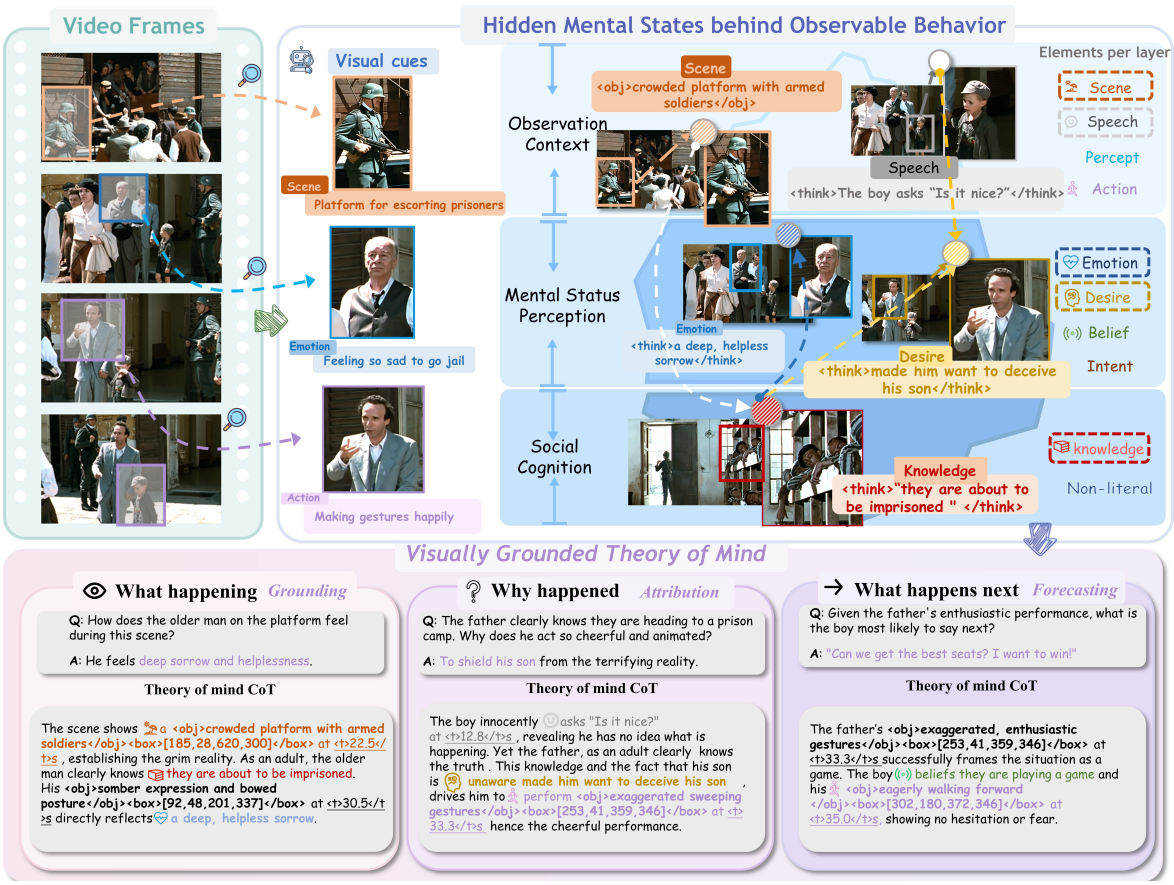

Figure 1: Overview of ToM-Video Bench. Each video is represented as a grounded Theory-of-Mind graph that links observable cues, such as scenes, speech, and actions, to latent mental states, including emotions, desires, beliefs, perception, intentions, knowledge, and non-literal communication. The resulting annotations support multiple question types, including evidence grounding, mental-state attribution, and future-state forecasting.

Once both streams are flattened into a single textual description, such cross-modal conflict can no longer be explicitly recovered.

To fill this gap, we introduce **ToM-Video**, a realistic video benchmark that jointly annotates physical evidence and mental-state inferences with fine-grained grounding. As shown in Figure 1, ToM-Video organizes each clip as a structured graph in which observable cues such as actions, speech, and perceptual evidence are linked to higher-level states such as beliefs, intentions, and emotions through modality-specific grounding. Visual nodes are anchored to timestamps and, when directly observable, keyframe-level bounding boxes and inference nodes are linked to their supporting premises. Based on these annotations, we construct question sets that probe complementary aspects of video ToM reasoning, including future prediction, backward causal inference, and evidence grounding. ToM-Video contains 41,206 multiple-choice questions over 1,487 referenced videos with movie-disjoint splits. Our benchmark moves multimodal ToM from a text-mediated reasoning task to one that is **visually trainable** , turning the temporally grounded visual stream into a first-class supervisory signal instead of a mere precursor to text.

Beyond evaluation, ToM-Video also enables benchmark-specific post-training. Its annotations supervise not only the final answer, but also the temporal, spatial, and source-chain evidence that supports each mental-state inference. We therefore develop **INSPIRE**, a staged framework that learns structured evidence traces through SFT, improves spatio-temporal grounding through RL-Grounding, and aligns grounded traces with ToM-oriented answer selection through RL-ToM. This curriculum separates format learning, evidence localization, and mental-state decision-making, avoiding the instability of optimizing all process rewards from the outset.

In summary, our contributions are as follows:

- We introduce **ToM-Video**, a large-scale real-video benchmark that makes ToM rea-

soning visually trainable by using temporally grounded video evidence as supervision, shifting multimodal ToM from text-mediated social inference to evidence-to-mind learning.

- To enable MLLMs to perform ToM perception in real video scenarios, we develop **INSPIRE**, a staged post-training framework that turns ToM-Video annotations into process rewards, decomposing video ToM learning into structured trace learning, spatio-temporal grounding, and RL-ToM optimization.

- We conduct comprehensive experiments showing that current MLLMs remain weak at transition-sensitive ToM grounding, while INSPIRE improves answer accuracy and produces more compact evidence-backed reasoning on both ToM-Video and external social-reasoning video benchmarks.

2 Related Work

Theory of Mind Benchmarks. Existing ToM benchmarks span text-only, synthetic-video, and real-video settings. Text-only benchmarks such as ToMi (Le et al., 2019) and HI-TOM (Wu et al., 2023) focus on nested belief tracking and logical inference through short stories. Synthetic video benchmarks such as MMToM-QA (Jin et al., 2024) and MuMA-ToM (Shi et al., 2025) evaluate goal and belief inference in controlled single- or multi-agent environments, but lack the social complexity and visual diversity of real scenes. Recent real-video efforts mark significant progress. MoMentS (Villa-Cueva et al., 2025) evaluates seven ToM categories over narrative-rich short films with 2.3K questions, EgoToM (Li et al., 2025b) benchmarks goal, belief, and action prediction from egocentric video, and ToM-SSI (Bortoletto et al., 2025) tests situated social interactions with up to four agents. SoMi-ToM (Fan et al., 2026) further studies multi-perspective ToM in embodied social interactions, emphasizing perspective-taking under temporal and social context. Video-Only ToM (Liu et al., 2026) instead focuses on improving MLLMs' ToM reasoning from video-only inputs, highlighting the importance of reducing language-prior reliance. VToM (Chen et al., 2025b) introduces frame localization for video-based ToM on Social-IQ 2.0, but its retrieval granularity is too coarse and has not reached the frame level. StoryTR (Zhong et al., 2026) also connects ToM-style reasoning with narrative-centric temporal retrieval. Unlike retrieval-oriented formulations, ToM-Video evaluates video question answering with explicit node-level ToM graphs and grounded reasoning chains. Closest to our work, R^3-VQA (Niu et al., 2026) annotates mental-state causal chains with an automated pipeline over movie clips. However, these benchmarks either lack keyframe-level spatial grounding or do not provide unified node-level ToM graphs with timestamped evidence, bounding boxes, explicit physical–mental links, and grounded chain supervision.

Video Spatio-Temporal Grounding and Training. Recent work has improved video grounding by explicitly localizing question-relevant evidence in time and space. Open-o3 Video (Meng et al., 2026) trains models to generate grounded reasoning traces with <t>, <obj>, and <box> tokens, while VisionCoach (Lee et al., 2026) improves grounding through visual prompting and object-aware rewards. Video-R1 (Feng et al., 2025) and Video-TwG (Chen et al., 2026) further explore GRPO-style or on-policy optimization for video reasoning and temporal grounding. However, these methods primarily focus on physical events and spatio-temporal evidence, without explicitly modeling the mental states behind human behavior. Our framework extends grounded video training from physical event reasoning to video ToM reasoning, using a three-stage curriculum with process-aware rewards to align grounded visual evidence with mental-state inference.

3 ToM-Video Benchmark Construction

3.1 Overview

ToM-Video is designed to expose the evidence behind each video ToM inference. Unlike conventional VideoQA, where the annotated target is often an event or object answer, our annotation target is the evidence path from observable cues to latent mental states.

As shown in Figure 2, the benchmark is constructed through six stages: video filtering, ToM-node extraction, visual evidence verification, keyframe-level spatial grounding, grounded QA generation, and quality control. Our core annotations are, first, nodes, used to describe multimodal evidence nodes in the video, including speech, visual actions, belief and emotion reasoning, etc., and

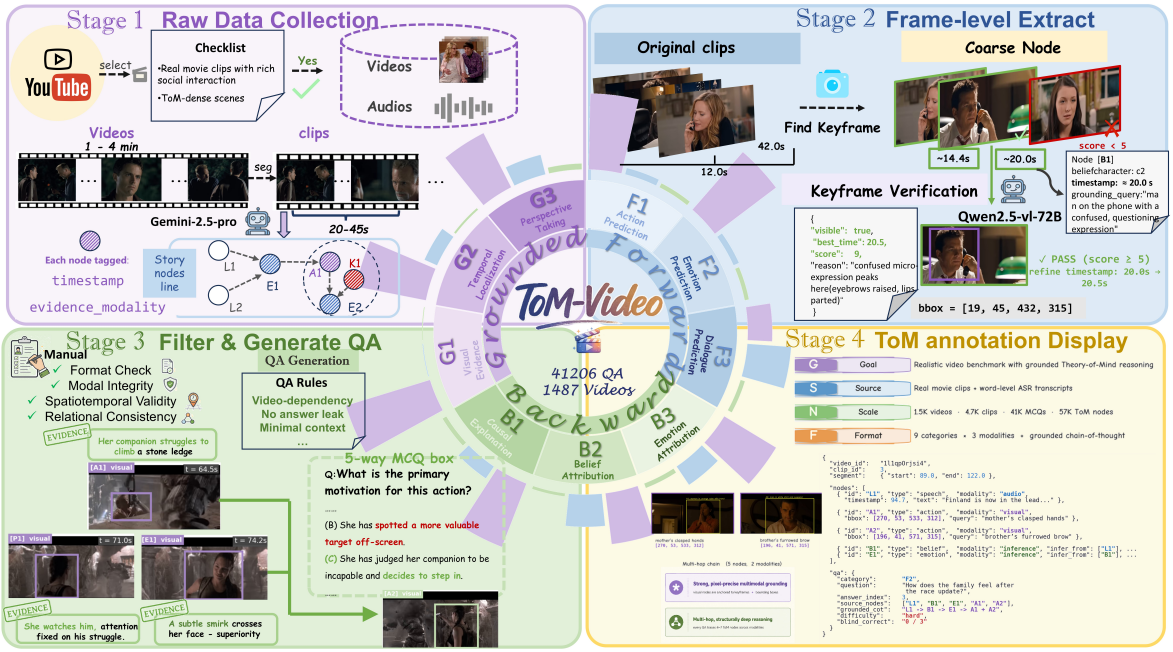

Figure 2: ToM-Video benchmark construction pipeline.

record their timestamps, bounding boxes, or dependencies; second, QA, used to store questions, answer indices, question categories, source evidence nodes, and reasoning chains.

3.2 Data Sources and ToM Ontology

Video collection. We collect candidate videos from YouTube and use an LLM-assisted filtering procedure to select realistic movie clips with rich social interactions and observable cues for mental-state inference. The filter removes clips that are primarily action-driven, visually ambiguous, dialogue-only, or lack meaningful interpersonal dynamics. The retained videos are processed into aligned video frames, audio, and ASR transcripts. After manually ensuring the validity of the videos, the final pool contains 1,487 source videos. We used Gemini-2.5-Pro to segment the ToM dense clips in each original video, ultimately dividing them into 4,666 clips which across diverse social contexts, including conflict, family, and outdoor scenarios. Detailed category statistics are provided in Appendix B.

ToM ontology. We define a three-layer ontology, details can be found in Appendix A. Layer 1 contains observable evidence, such as scene descriptions, actions, dialogue, and perceptual access. Layer 2 captures latent mental states, including beliefs, intentions, desires, and emotions. Layer 3 represents higher-order social-cognitive relations, such as knowledge and non-literal meaning. Each node is assigned an evidence modality: *visual*, *audio*, or *inference*. Visual nodes require temporal and spatial grounding, audio nodes require ASR alignment, and inference nodes must cite supporting lower-level nodes.

3.3 Annotation Pipeline

ToM-node extraction. We first process long videos with a sliding-window strategy. For each window, Gemini-2.5-Pro analyzes the video together with the ASR transcript to identify **ToM-dense** interaction segments, such as moments where a character's belief, intention, emotion, or knowledge state changes in response to observable evidence. Each selected segment is converted into a short clip and annotated with structured ToM nodes, including node type, character, timestamp, evidence modality, description, and causal links. For visual nodes, the model also generates a generic grounding query that describes the visible entity or region without relying on character names.

Evidence verification and grounding. Because automatically extracted visual claims may be temporally imprecise or not directly grounded in observable evidence, we introduce a two-stage verification and calibration mechanism. First, we re-examine each visual node within a short temporal neighborhood around its predicted timestamp using Qwen-2.5-VL-72B to ensure that the described cue

is actually visible and to refine its temporal alignment. This step explicitly filters out hallucinated or over-inferred visual claims that cannot be localized in the video.

Second, we perform spatial grounding using a local Qwen3-VL-8B model, which predicts bounding boxes for the corresponding keyframes based on the generated grounding query. This design separates temporal validation from spatial localization, allowing us to reduce error propagation across modalities. Boxes that are degenerate, off-canvas, or inconsistent with the visual evidence are discarded, ensuring that only visually verifiable signals enter the final graph.

QA generation. Given this structured graph, we design a **nine-category QA taxonomy** that systematically covers different directions of ToM reasoning, including forward prediction, backward explanation, and grounded perception. This design ensures that questions are not limited to answer recognition, but explicitly test whether models can (i) predict behavioral consequences from mental states, (ii) infer latent causes of observed actions, and (iii) localize the spatio-temporal evidence supporting a ToM inference. Gemini-2.5-Pro generates five-way multiple-choice questions with plausible distractors, question-relevant source nodes, and grounded reasoning traces containing timestamps, object mentions, and bounding boxes when available. Details are in Appendix D.

Automatic filtering. Before human review, we apply automatic checks to remove hallucinated visual nodes, invalid boxes, incomplete source-node chains, ambiguous QAs, questions answerable from text alone, and samples lacking sufficient video evidence. Further details of automatic filtering and human verification are provided in Appendix E.

3.4 Human Quality Control

We also conduct human verification because ToM reasoning often depends on subtle social, perceptual, and pragmatic judgments. Ten trained reviewers with backgrounds in cognitive science and AI participate in the audit. Each sample is independently reviewed by two annotators, who verify whether visual nodes are perceivable at the cited timestamp, boxes cover the intended entity, the question has exactly one correct answer without requiring external knowledge. The Appendix E.2 shows the details.

4 INSPIRE: Process-Aware Training for Grounded Video ToM Reasoning

To enhance the model's ability to find evidence of ToM from videos, we introduce INSPIRE (short for INtention-aware Spatio-temporal and Perspective-Informed REwarding), a three-phase curriculum that progressively teaches a model to produce structurally valid reasoning traces, ground these traces in spatio-temporal evidence, and align grounded evidence with Theory-of-Mind inference and final answer selection. Given a model response y, we denote by τ the reasoning trace enclosed in <think>...</think> and by o the final answer option enclosed in <answer>...</answer>.

4.1 Process Rewards

A grounded ToM response should be parseable, grounded in the relevant video evidence, ToM-relevant, and consistent with the annotated source chain. We therefore define four process rewards. Given a response $y = (\tau, o)$, we parse a set of predicted evidence units $\hat{\mathcal{E}}$ from τ. Each unit contains a predicted timestamp t, an optional bounding box b, and an optional source-node reference.

R_{str}: Format reward. Malformed traces cannot be reliably grounded or scored. We define a type-specific required tag sequence $\mathcal{S}_{\text{req}} = (s_1, \ldots, s_N)$. For grounding questions, this sequence contains <think>, temporal tags <t>, object tags <obj>, bounding-box tags <box>, and <answer>; for other question types, the box tags are omitted. A tag is counted only if its content is parseable, e.g., timestamps must be numeric and boxes must contain four valid coordinates when required. The reward is the fraction of required tags that appear in the correct order:

$$R_{\text{str}}(y) = \frac{1}{N} \sum_{i=1}^{N} v_i, \tag{1}$$

where $v_i = 1$ if s_i appears after all previously matched tags and is parseable, and $v_i = 0$ otherwise.

R_{grd}: Spatio-temporal grounding. Let $\mathcal{E}^\star$ be the annotated evidence units for question q. For a predicted unit $\hat{e}_i = (\hat{t}_i, \hat{b}_i)$ in $\hat{\mathcal{E}}$ ($\hat{t}_i$ represents the time when the prediction node occurs and $\hat{b}_i$ represents the predicted bounding box of the object), a gold unit $e_j^\star = (t_j^\star, b_j^\star)$, we compute a soft grounding score s_{ij}. The temporal window

is $\Delta(\gamma) = \max(1.0, 4.0(1-\gamma))$, where $\gamma \in [0,1]$ denotes normalized training progress, increasing from 0 to 1 over the RL stage. The window is loose at the beginning and becomes stricter later, so early approximate timestamps can still receive reward while the model is gradually pushed toward precise temporal grounding. If $|\hat{t}_i - t_j^\star| \geq \Delta(\gamma)$, then $s_{ij} = 0$; otherwise,

$$s_{ij} = \exp\left(-\frac{(\hat{t}_i - t_j^\star)^2}{2(\Delta(\gamma)/2)^2}\right) \cdot \beta_{ij}, \qquad (2)$$

where $\beta_{ij} = \mathrm{IoU}(\hat{b}_i, b_j^\star)$ for visual evidence with boxes, and $\beta_{ij} = 1$ otherwise. We perform one-to-one matching $\mathcal{M}$ between predicted and gold evidence units by maximizing s_{ij}, and define

$$R_{\text{grd}}(y) = \frac{1}{2}\left(\frac{\sum_{(i,j)\in\mathcal{M}} s_{ij}}{\max(1, |\hat{\mathcal{E}}|)} + \frac{\sum_{(i,j)\in\mathcal{M}} s_{ij}}{\max(1, |\mathcal{E}^\star|)}\right). \qquad (3)$$

This penalizes both irrelevant extra evidence and missing gold evidence.

R_{cnt}: Answer-conditioned ToM support. A trace should not receive high ToM-process reward if it supports an incorrect answer. We therefore gate ToM trace quality by final-answer correctness:

$$R_{\text{cnt}}(y) = \mathbb{1}[o = o^\star]\big(\lambda_a + \lambda_s \rho_s(\tau, q) + \lambda_p \rho_p(\tau)\big). \qquad (4)$$

Here, ρ_s measures agreement between the source-node references claimed in the trace and the annotated source nodes, while ρ_p rewards explicit perspective evidence when it is tied to a cited evidence unit. We set $(\lambda_a, \lambda_s, \lambda_p) = (0.5, 0.3, 0.2)$, where λ_a gives base credit for the correct answer, λ_s rewards source-node alignment, and λ_p rewards perspective-aware evidence. Details are in Appendix H.

R_{cns}: Evidence-chain consistency. A correct answer may still fail to use logical ToM evidence. We therefore reward traces that cite source nodes in the same order as the gold ToM chain. Let $\mathcal{C}^\star = ((r_1^\star, t_1^\star, \ell_1), \ldots, (r_m^\star, t_m^\star, \ell_m))$ be the ordered gold chain, where $r_j^\star$ is the ToM node, $t_j^\star$ is the timestamp, and ℓ_j is the ontology layer. Let $\hat{\mathcal{C}} = (\hat{r}_1, \hat{t}_1), \ldots, (\hat{r}_M, \hat{t}_M)$ be the ordered node claims parsed from τ. A gold node j is matched if there exists a predicted claim i such that $\hat{r}_i = r_j^\star$ and $|\hat{t}_i - t_j^\star| < \delta$, where $\delta = 4\,\mathrm{s}$. Let $\pi(j)$ be the first matched position of node j in $\hat{\mathcal{C}}$. We define

$$R_{\text{cns}}(y) = g(y) \cdot \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}[\pi(j) \neq \varnothing] \cdot \phi_j, \qquad (5)$$

where

$$\phi_j = \begin{cases} 1, & \text{if } j = 1 \text{ or } \pi(j) \geq \pi(j-1), \\ \eta, & \text{otherwise.} \end{cases} \qquad (6)$$

We set the order-discount factor $\eta = 0.5$. The consistency gate $g(y)$ is set to 0 if the trace explicitly supports an answer option different from the final answer o, and to 1 otherwise.

4.2 Three-Phase Curriculum Training

Jointly maximizing all reward signals from the start leads to training instability. We therefore adopt a three-phase curriculum.

Phase 1: Supervised fine-tuning. A randomly initialized policy almost never emits a syntactically valid trace, so We warm-start π_θ on a curated subset $\mathcal{D}_{\text{SFT}} \subset$ ToM-Video, in which each example $(x, y^\star)$ pairs the video–question input with a gold structured response $y^\star = (\tau^\star, o^\star)$ whose trace $\tau^\star$ is materialized from our ontology annotations and already contains the <think>, temporal, object, box, and <answer> tags expected by all four process rewards. We minimize the standard token-level negative log-likelihood on the assistant tokens,

$$\mathcal{L}_{\text{SFT}}(\theta) = -\,\mathbb{E}_{(x,y^\star)\sim\mathcal{D}_{\text{SFT}}} \sum_t \log \pi_\theta\big(y_t^\star \mid x, y_{<t}^\star\big). \qquad (7)$$

Phase 2: RL-Grounding. Starting from the SFT checkpoint, we apply GRPO (Shao et al., 2024) on the subset of ToM-Video with explicit grounding supervision. The reward in this phase is

$$R^{(2)}(y) = \mu_{\text{str}}\, R_{\text{str}}(y) + \mu_{\text{grd}}\, R_{\text{grd}}(y), \qquad (8)$$

with $(\mu_{\text{str}}, \mu_{\text{grd}}) = (1.0,\ 1.5)$. We upweight R_{grd} because R_{str} saturates within the first few hundred steps; the exact ratio is fixed by a small grid search. Both rewards are computed against freely sampled completions. We adopt GSPO-style sequence-level importance weighting (Zheng et al., 2025) and DAPO-style asymmetric clipping (Yu et al., 2025); full settings are deferred to Appendix H.

Phase 3: RL-ToM. Starting from the RL-Grounding checkpoint, this phase replaces the active reward set with $\{\text{grd}, \text{cnt}, \text{cns}\}$:

$$R^{(3)}(y) = \mu_{\text{grd}}\, R_{\text{grd}}(y) + \mu_{\text{cnt}}\, R_{\text{cnt}}(y) + \mu_{\text{cns}}\, R_{\text{cns}}(y), \tag{9}$$

with $(\mu_{\text{grd}}, \mu_{\text{cnt}}, \mu_{\text{cns}}) = (0.3,\ 1.0,\ 1.0)$. We drop R_{str} because format compliance is already near ceiling after SFT and Phase 2. The chosen weighting preserves a weak grounding constraint while shifting most of the reward mass to answer quality and evidence-trace consistency. We again use GRPO, with full optimization settings deferred to Appendix H.

5 Experiments

5.1 Experimental Setup

We evaluate all models on TOM-VIDEO-VAL, which contains 2,162 multiple-choice questions.

Baselines. We compare INSPIRE with a broad set of closed-source and open-source video-language models. The closed-source group includes GPT-4o, GPT-5.4-mini, GPT-5.4, Gemini-2.5-Flash, Gemini-3-Flash, and Claude Sonnet 4.6. The open-source group includes general-purpose and reasoning-oriented video-LLMs from 2B to 235B parameters.

Evaluation protocol. To keep the visual input budget comparable, all vision-language models are evaluated with 32 uniformly sampled frames resized to at most 448 pixels on the longer side. No annotated keyframes, bounding boxes, source nodes, or reasoning chains are provided at test time. All models are prompted in the same multiple-choice format.

5.2 Experimental Results

> ***Finding 1:*** Strong video-language models still struggle with transition-sensitive ToM grounding.

Table 2 shows that INSPIRE attains 76.2% accuracy, the best among all evaluated systems. Scaling alone is insufficient: Qwen3-VL improves from 43.4% at 2B to 71.2% at 235B, still below the trained 8B model. The hardest residual category is temporal grounding (G2), where even INSPIRE reaches only 50.3% versus 92.5% human, indicating that locating when a mental-state shift occurs remains a major bottleneck.

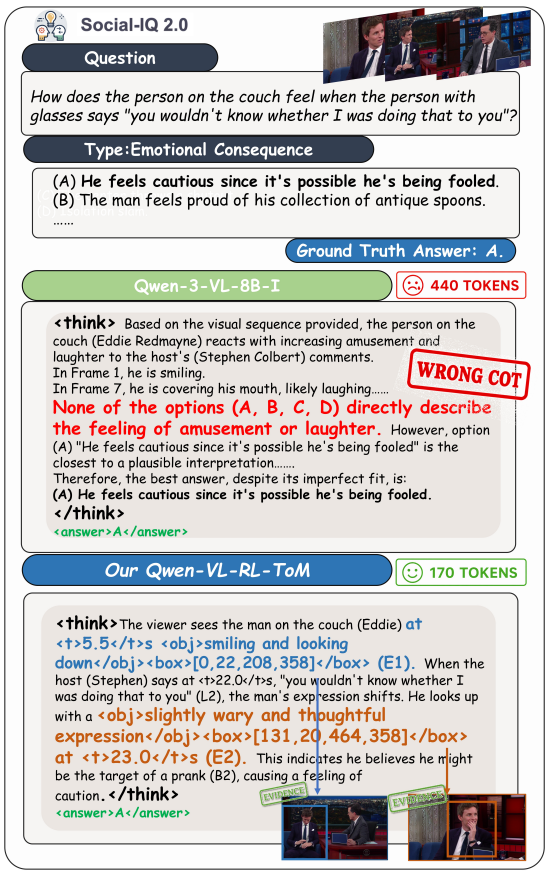

Figure 3: A shared-correct Social-IQ example.

The staged curriculum separates grounding from ToM decision-making. SFT learns the structured trace format; RL-Grounding improves evidence localization with little answer-accuracy gain (70.8% to 71.0%); and RL-ToM raises accuracy to 76.2%.

External generalization. We additionally evaluate Qwen3-RL-ToM on external video and social-reasoning benchmarks, including Video-MME (Fu et al., 2025) and Social-IQ 2.0 (Wilf et al., 2023). The model improves over the Qwen3-VL-8B backbone on both benchmarks, indicating that the learned grounded ToM behavior transfers beyond TOM-VIDEO. Our trained model increased the accuracy of video-MME short videos from 73.16% to 75.76%, and improved the accuracy of Social-IQ from 63.41% to 65.12%. Details are shown in Appendix K.

> ***Finding 2:*** Grounded ToM training reallocates reasoning from surface description to compact evidence-backed mental-state inference.

To test our trained model's reasoning quality, we compare the two models on the shared-correct sub-

	Model	Backward			Forward			Grounded			Acc
		B1	B2	B3	F1	F2	F3	G1	G2	G3	
Closed	GPT-4o (Hurst et al., 2024)	84.1	73.9	76.9	77.7	72.6	66.0	82.0	28.8	81.2	66.4
	Gemini-2.5-Flash (Comanici et al., 2025)	67.1	59.6	59.8	55.7	71.6	57.8	71.2	36.1	64.2	56.8
	Gemini-3-Flash (Google DeepMind, 2025)	89.9	81.2	79.9	79.4	88.4	80.9	97.3	39.5	83.9	74.7
	GPT5.4-mini (Singh et al., 2025)	80.2	72.0	69.4	74.1	66.3	53.1	90.1	37.3	80.7	64.6
	GPT5.4 (Singh et al., 2025)	91.8	80.7	84.3	81.6	81.1	76.6	91.9	36.1	92.2	74.2
	Claude Sonnet 4.6	89.9	75.7	85.2	83.0	85.8	84.4	92.8	37.9	89.9	75.3
Open	InternVL-2.5-8B (Chen et al., 2024a)	63.8	63.8	58.1	46.8	68.9	46.1	87.4	39.0	60.6	55.0
	LLaVA-OV-7B (Li et al., 2024)	63.8	60.1	57.6	58.5	69.5	45.7	82.9	42.1	61.0	56.6
	Video-R1-7B (Feng et al., 2025)	56.0	58.7	53.3	49.3	64.2	35.9	58.6	38.6	55.0	49.9
	VideoChat-R1 (Li et al., 2025a)	74.9	71.1	75.1	61.0	78.9	54.3	80.2	39.2	69.7	63.0
	Qwen2.5-VL-7B (Bai et al., 2025b)	60.4	55.0	69.9	51.1	68.9	46.9	73.9	36.4	56.9	54.1
	Qwen3-VL-2B-I (Bai et al., 2025a)	50.7	49.1	40.2	42.2	54.7	44.1	64.9	28.8	44.5	43.4
	Qwen3-VL-4B-I (Bai et al., 2025a)	75.8	67.4	67.7	69.9	80.0	55.1	74.8	38.4	70.2	62.8
	Qwen3-VL-8B-I (Bai et al., 2025a)	72.0	60.6	59.0	53.5	58.4	53.9	71.2	33.7	64.2	54.9
	Qwen3-VL-8B-T (Bai et al., 2025a)	78.7	67.4	72.5	60.3	70.5	57.8	91.9	24.8	74.3	60.3
	Qwen3-VL-30B-A3B-I (Bai et al., 2025a)	86.5	72.0	76.9	60.6	70.0	57.8	85.6	33.9	80.7	64.2
	Qwen3-VL-30B-A3B-T (Bai et al., 2025a)	**91.3**	75.7	72.9	69.1	81.1	57.8	89.2	32.8	84.4	67.0
	Qwen3-VL-235B-A22B-T (Bai et al., 2025a)	90.3	79.8	80.8	71.6	78.9	72.3	**94.6**	36.1	86.7	71.2
	Qwen3-VL-8B + SFT	79.7	81.7	77.3	75.9	73.7	74.6	84.7	41.2	84.9	70.8
	Qwen3-VL-8B + RL-Grounding	81.2	78.9	80.3	71.6	76.8	72.3	83.8	44.6	84.9	71.0
	Qwen3-VL-8B + RL-ToM	86.0	**82.6**	**83.0**	**77.3**	**84.2**	**80.1**	87.4	**50.3**	**88.1**	**76.2**
	Human Level	95.2	91.3	93.0	92.9	93.7	96.5	92.8	92.5	96.8	**93.8**

Table 2: Main results on ToM-Video-val (2,162 multiple-choice questions). B/F/G denote *Backward*, *Forward*, and *Grounded* reasoning categories; **Acc** is overall accuracy (%). Best and second-best are highlighted independently within *closed* (best / 2nd) and *open* $\cup$ *ours* (best / 2nd); the two groups are ranked separately.

Dataset	Model	$N_\cap$	$\bar{\tau}\downarrow$	$\bar{\epsilon}\uparrow$	$\bar{\chi}\uparrow$
Social-IQ	Base	273	215.9	1.13	60.3
Social-IQ	RL-ToM	273	115.7	3.98	73.6
Ours-VAL	Base	535	131.7	0.32	59.4
Ours-VAL	RL-ToM	535	124.2	4.14	81.2

Table 3: Process-quality statistics on correct subsets. $N_\cap$ is the shared-correct subset Q nums; $\bar{\tau}$ **is average reasoning length in tokens**; $\bar{\epsilon}$ is the average number of explicit evidence mentions (timestamps, boxes, or source references); $\bar{\chi}$ is human chain validity from double-blind annotation (Appendix K.4).

set, where both models select the correct answer. Table 3 shows that RL-ToM uses fewer reasoning tokens while producing more explicit evidence mentions and higher human-rated chain validity. On Social-IQ, RL-ToM reduces the average reasoning length from 215.9 to 115.7 tokens, while human chain validity from 60.3 to 73.6. A similar pattern appears on ToM-Video-val, where RL-ToM yields denser evidence traces and higher chain validity despite a shorter average trace. This suggests that the gain is not simply longer CoT, but more selective use of ToM-relevant evidence.

Figure 3 illustrates this difference. The base model correctly selects option A, but its rationale mainly enumerates visible laughter and treats the answer as the closest plausible option. RL-ToM instead grounds the inference in two visual moments: the guest's initial expression at 5.5s and the later wary expression at 23.0s after the host's utterance.

6 Conclusion

We introduced ToM-Video, an evidence-grounded benchmark for Theory-of-Mind reasoning in realistic videos, and **INSPIRE**, a staged post-training framework for ToM training. More broadly, this work studies mental-state reasoning under passive observation. A natural next step is to connect it with *Theory of Space* (Zhang et al., 2026), agents should not only infer beliefs and intentions from observed evidence, but also actively decide what to observe next, reducing uncertainty and updating spatial and mental beliefs through sequential interaction.

Limitations

ToM-Video focuses primarily on short, interaction-dense clips, which makes fine-grained timestamp, bounding-box, and source-chain annotation feasible for video training and inference. Longer videos may contain more distributed social cues and delayed mental-state changes, so extending grounded ToM reasoning to long-form scenarios remains future work.

Our analysis focuses on observable outputs and annotation-aligned evidence traces, rather than on the model's internal mechanisms. We do not identify which parameters, layers, attention heads, or hidden representations are most sensitive to ToM-relevant variables. Therefore, the results should be interpreted as evidence of improved external ToM grounding, not as a mechanistic account of how ToM is represented inside the model. Future work could combine ToM-Video with probing, activation patching, and component-level ablations to localize ToM-sensitive representations and circuits.

Ethics Statement

ToM-Video is intended to support research on evidence-grounded video Theory-of-Mind reasoning for human-centered AI systems. The benchmark evaluates whether models can link observable visual and auditory cues to annotated mental-state inferences, rather than infer the private mental states of real individuals. We explicitly discourage downstream use for surveillance, behavioural profiling, hiring or credit-scoring decisions, clinical diagnosis, or any consequential judgement about identifiable people.

All source clips are sampled from publicly available, narrative movie content uploaded to YouTube. We do not crawl private or restricted-access content. Because the YouTube Terms of Service do not allow redistribution of the underlying video files, our release contains only (i) YouTube video IDs, (ii) clip start/end timestamps, (iii) ASR transcripts produced by us, and (iv) our annotations (ToM graphs, QA pairs, grounded chains-of-thought, and bounding boxes). Users reconstruct the local video files from the IDs through standard YouTube tooling and remain bound by YouTube's Terms of Service when doing so. Our textual annotations are released under CC BY-NC 4.0 for non-commercial research.The ten human reviewers were graduate-level researchers with backgrounds in cognitive science or AI. They provided informed consent, were compensated above the local minimum hourly wage, could opt out at any time, and were given a content-warning protocol with the option to skip distressing clips. No personal data of the annotators is released. Mental-state labels are treated as interpretive judgements grounded in observable cues, not as ground truth about any real person's mind.

Upon acceptance, we will release the annotation files, metadata, prompts, training, and evaluation code through an anonymized repository.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv: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, and 8 others. 2025b. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064.

Simon Baron-Cohen. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, Cambridge, MA.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. 2025. ToM-SSI: Evaluating theory of mind in situated social interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32264–32289, Suzhou, China. Association for Computational Linguistics.

Tania Chakraborty, Eylon Caplan, and Dan Goldwasser. 2025. VIBE: Can a VLM read the room? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22992–23008, Suzhou, China. Association for Computational Linguistics.

Houlun Chen, Xin Wang, Guangyao Li, Yuwei Zhou, Yihang Chen, Jia Jia, and Wenwu Zhu. 2026. Think with grounding: Curriculum reinforced reasoning with video grounding for long video understanding. *ArXiv*, abs/2602.18702.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025a. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria. Association for Computational Linguistics.

Zhanwen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2025b. Through the theory of mind's eye: Reading minds with multimodal video large language models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. 2026. Somitom: Evaluating multi-perspective theory of mind in embodied social interactions. *Advances in Neural Information Processing Systems*, 38.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Google DeepMind. 2025. Gemini 3 pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Model card update: December 2025; model release: November 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Julian Jara-Ettinger. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. 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)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. 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)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. 2026. Visioncoach: Reinforcing grounded video reasoning via visual-perception prompting. *Preprint*, arXiv:2603.14659.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025a. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Yuxuan Li, Vijay Veerabadran, Michael L Iuzzolino, Brett D Roads, Asli Celikyilmaz, and Karl Ridgeway. 2025b. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152*.

Siqi Liu, Xinyang Li, Bochao Zou, Junbao Zhuo, Huimin Ma, and Jiansheng Chen. 2026. Video-only

tom: Enhancing theory of mind in multimodal large language models. *arXiv preprint arXiv:2603.24484*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. 2026. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *Preprint*, arXiv:2510.20579.

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

Josef Perner. 1991. *Understanding the representational mind.* The MIT Press.

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

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

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

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1510–1519.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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

Henry M Wellman, Susan Carey, Lila Gleitman, Elissa L Newport, and Elizabeth S Spelke. 1990. *The child's theory of mind*. The MIT Press.

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

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, and 1 others. 2026. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *The Fourteenth International Conference on Learning Representations*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Xuanyue Zhong, Yuqiang Xie, Guanqun Bi, Jiangping Yang, and Guibin Chen. 2026. Storytr: Narrative-centric video temporal retrieval with theory of mind reasoning. *arXiv preprint arXiv:2604.23198*.

Overview. This appendix provides additional details and analyzes to complement the main paper.

- Section A defines the mental-state node types used in the ToM reasoning traces.
- Section B reports the category distribution of the video pool and analyzes dataset coverage.
- Section C presents the benchmark construction prompts used for video filtering, node extraction, visual verification, bounding-box annotation, and QA generation.
- Section D describes the design of the nine QA categories and their reasoning targets.
- Section E details the automatic filtering and human verification procedures used for quality control.
- Section F provides additional statistics on corpus scale, ontology coverage, surface properties, and category-level QA counts.
- Section G reports the training configuration for SFT, RL-Grounding, and RL-ToM.
- Section H gives the implementation details of the reward functions and GRPO-based optimization.
- Section I describes the movie-level train/validation split and data leakage checks.
- Section J evaluates the faithfulness of intermediate grounding traces and presents input-side and reward-side ablations on ToM-Video-val.
- Section K reports additional external generalization results on Video-MME and Social-IQ 2.0.
- Section L provides qualitative case studies comparing the baseline and the trained ToM model.
- Section M discloses the use of large language models in annotation, training, evaluation, and writing assistance.

A Mental-State Node Definitions

A world model not only requires a physical model component, but we also need to make objects useful to humans, which is the mental model. These two parts support each other in order to achieve a world model that can interact with humans. Elements of the theory of mind are an important part of constructing a mental world model, and we provide the following explanations for the elements used in the paper.

The ontology is designed to cover seven canonical elements of Theory of Mind: beliefs, intentions, emotions, perception, desires, knowledge, and non-literal meaning. Since these elements are not always directly observable in video, we ground them through a 10-node annotation scheme that links latent mental-state and social-cognitive variables to perceptual evidence. We define ten ToM node types $n_k \in \{D, A, L, P, B, I, R, E, K, N\}$, organized into a three-layer ontology. Layer 1 contains *observable evidence*: scene descriptions, actions, dialogue, and perceptual access. Layer 2 contains *latent mental states*: beliefs, intentions, desires, and emotions, which are not directly observed but inferred from Layer-1 evidence. Layer 3 contains *higher-order social-cognitive processes*: knowledge asymmetry and non-literal communication, which require reasoning over multiple agents' information states, speaker intentions, and pragmatic context. This hierarchy follows the standard belief–desire view of action explanation, where beliefs and desires jointly support intentions and actions (Wellman et al., 1990; Perner, 1991), and extends it with perspective-taking and pragmatic inference for multi-agent social interaction (Baron-Cohen, 1995; Chen et al., 2025a). Table 4 summarizes the node types; below we describe the annotation rule for each type.

D — Description. Static scene context: spatial layout, salient objects, lighting, and social setting (e.g., *hospital waiting room*, *office cubicle*). The environment constrains which mental states are plausible—a hospital and a party prime very different inferences—so D nodes serve as the contextual prior for every downstream Layer-2 inference.

A — Action. Intentional physical behaviour: gestures, posture shifts, gaze direction, approach/avoidance movements, and *micro-actions* such as lip-biting, fidgeting, or gaze-aversion. In Wellman's belief–desire framework, actions are

Sym.	Name	Layer	Modality
D	Description	L1 (input)	visual
A	Action	L1 (input)	visual
L	Dialogue	L1 (input)	audio
P	Percept	L1 (input)	visual
B	Belief	L2 (mental)	inference
I	Intent	L2 (mental)	inference
R	Desire	L2 (mental)	inference
E	Emotion	L2 (mental)	visual / inf.
K	Knowledge	L3 (social)	inference
N	Non-literal	L3 (social)	audio / inf.

Table 4: Ten mental-state node types referenced by reasoning-trace markers n_k. Modality indicates the typical evidence source for the node.

the primary observable output of the belief–desire system; ToM annotations reverse-engineer mental states from what people *do*.

L — Dialogue. Speech transcribed from ASR. Treated as a dual-channel signal: *content* (propositional meaning) and *manner* (tone, hesitation, volume, speed). L nodes are anchored by transcript timestamps and speaker localization rather than bounding boxes.

P — Percept. What a character perceives—gaze direction, attentional focus, overhearing, or noticing something others miss. Following Baron-Cohen's Shared Attention Mechanism (Baron-Cohen, 1995), P nodes determine each character's information state and are therefore prerequisites for false-belief and perspective-taking reasoning.

B — Belief. What a character takes to be true about the world, themselves, or others' minds. Annotated with explicit level: *1st-order* ("X believes P"), *2nd-order* ("X believes that Y believes P"), and *false belief* (the character's belief contradicts reality).

I — Intent. The character's goal, plan, or strategy. Three sub-levels are recognized: *proximal* (immediate action goal), *distal* (higher-level purpose), and *strategic* (deliberate manipulation of others' mental states). The depth of intent annotation is what separates ToM data from action-recognition data.

E — Emotion. Emotional states arising when events match or mismatch beliefs and desires. Each E node specifies (i) the observable cues—facial expression, vocal tone, body posture—and (ii) the appraisal that links those cues to a B or I node.

R — Desire. A character's wanted or preferred state of the world, such as wanting to leave, hoping to be believed, or wishing to hide information. In belief–desire accounts of folk psychology, desires specify what an agent wants, while beliefs specify what the agent takes to be true; together they help explain intentions and actions.

K — Knowledge. Background or world knowledge that some characters have and others lack. Knowledge asymmetry drives perspective-taking (G3) questions and is essential for false-belief scenarios; K nodes therefore carry an explicit list of *which characters share this knowledge*.

N — Non-literal. Communicative meaning that goes beyond the literal utterance, including irony, sarcasm, teasing, indirect requests and deception. We place non-literal meaning in Layer 3 because it requires reasoning over speaker intention, listener knowledge, shared context, and the gap between what is said and what is meant. An L node records the utterance itself; an N node records its pragmatic social meaning.

B Video categories

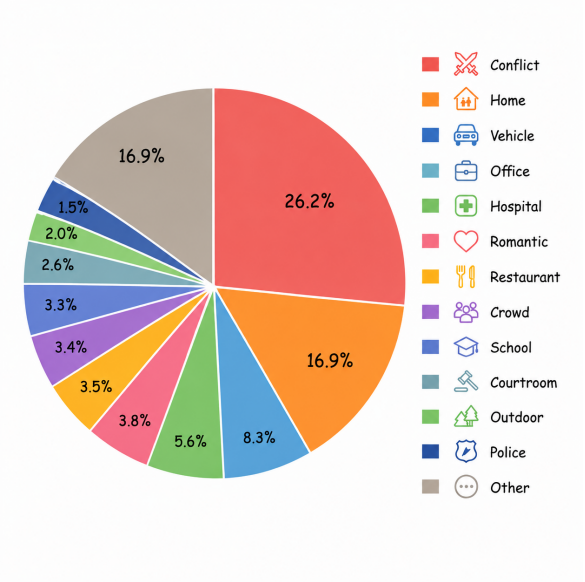

Figure 4: Distribution of our videos.

Figure 4 summarizes the category distribution of the ToM-Video video pool. The dataset contains 1,487 source videos and 4,666 clips across diverse social contexts. The largest category is *Conflict*

& Combat (26.2%), followed by *Home & Family* (16.9%) and *Other* (16.9%). These categories provide many ToM-relevant situations, including intention recognition, emotional reaction, belief updates, deception, and interpersonal misunderstanding.

The remaining videos cover travel, workplace, medical, romantic, dining, public-event, school, courtroom, outdoor, and police-related scenarios. This long-tailed distribution increases the diversity of social roles, environments, and interaction patterns, ensuring that ToM-Video does not rely on a single social script. Category labels are used only for analyzing dataset coverage; the final ToM graphs and QA pairs are grounded in clip-level visual evidence, audio, and source-node chains.

C Benchmark Construction Prompts

This appendix lists the core system prompts used by each stage of the ToM-Video construction pipeline (§3). Prompts are abridged for readability—verbose formatting instructions, output-schema examples, and few-shot blocks are replaced by [. . .]. The full prompts will be released with the pipeline code.

C.1 LLM-assisted Filtering for ToM-dense Videos

Before ToM annotation, we apply an LLM-assisted semantic filtering step using the GPT-120b-OSS model(Agarwal et al., 2025) to select clips suitable for grounded mental-state reasoning. To obtain reliable ASR and interaction-rich narratives, we focus on movie clips rather than unconstrained web videos. Movie clips typically provide clearer dialogue, coherent story context, and denser character interactions, which are important for annotating beliefs, intentions, emotions, knowledge states, and non-literal communication.

For each candidate movie clip, the model receives available metadata, caption-like visual summaries, and ASR transcript excerpts, and assigns a suitability score from 0 to 10. The score reflects whether the clip contains socially meaningful interaction, observable visual or auditory evidence, temporal progression, non-trivial mental-state ambiguity, and sufficient dialogue quality for ASR-based grounding.

The filtering output is used only as a candidate-level screening signal. High-scoring clips are retained for annotation, while low-scoring clips are removed. LLM-generated reasons are not treated as ground truth and are not copied into the final ToM graphs, question-answer pairs, or reasoning chains.

LLM Filtering System Prompt

```
You are helping select movie clips for a Theory−of−Mind video
        benchmark.

Your task is to judge whether a candidate movie clip is suitable for
        grounded
mental−state reasoning. We focus on movie clips because they usually
        contain
clear dialogue, coherent narrative context, and rich character
        interactions,
which are important for reliable ASR and fine−grained ToM
        annotation. A
suitable clip should contain observable social evidence that can support
inferences about what a person believes, wants, intends, feels, knows,
perceives, or implies.

Input:
− video_title: {video_title}
− visual_summary: {visual_summary}
− asr_excerpt: {asr_excerpt}
− duration: {duration}
− candidate_scene_type: {scene_type}

\textbf{(Omit)}

Reject the clip if:
− it is not a film/movie scene or movie−like narrative clip;
− it is a trailer, interview, commentary, recap, music video, or
        disconnected
  montage rather than a coherent scene;
− it has no human or human−like social agents;
− it is mainly a physical action scene without interpretable social
        cognition;
− it is dialogue−only with little or no visual grounding;
− the dialogue is too noisy or ASR quality is too poor for transcript
        alignment;
− the relevant evidence is too ambiguous or visually inaccessible;
− the clip cannot support keyframe−level evidence grounding.

Return ONLY a JSON object:
{
  "score": X,
  "suitable": true_or_false,
  "is_movie_clip": true_or_false,
  "asr_quality": "clear_usable_noisy_unusable",
  "narrative_coherence": "self_contained_partial_fragmented",
  "scene_type":
        "conflict_family_travel_workplace_medical_romantic_public
  _school_legal_outdoor_crime_other",
  "tom_cues": ["belief", "intent", "desire", "emotion", "knowledge",
        "perception", "non_literal"],
  "evidence_modalities": ["visual", "audio", "inference"],
  "reason": "brief explanation of why the clip is or is not suitable"
}
```

In our implementation, clips with score $\geq$ 7, suitable=true, is_movie_clip=true, and usable ASR quality are retained as annotation candidates.

C.2 VLM Node Extraction

Run on Gemini 2.5 Pro with the entire video window (45 s) uploaded inline alongside time-aligned ASR. The model returns a JSON document of clips, each with 8–15 structured nodes.

Prompt for VLM Node Extraction

```
You are the lead architect for a "Video Theory−of−Mind (ToM)"
        benchmark.
You are watching a REAL VIDEO (not still images). Pay close
        attention to
facial micro−expressions, body language shifts, gaze direction,
        gestures,
and tone changes across time.

### THE TASK
1. Scan the full video + ASR to find "Psychological Interaction
        Segments"
  where characters' mental states (beliefs, intentions, desires, emotions)
  are dense and evolving.
2. Slice each segment into a clip of {clip_min}{clip_max} seconds.
  IMPORTANT: Prefer LONGER clips (20−45 s) that capture a
        complete psychological
  arc (trigger  mental state change  behavioral consequence).
  Short clips (<20s) are REJECTED unless extremely ToM−dense.
3. Annotate each clip with 815 structured nodes (see below).
4. Each clip MUST have at least 5 nodes and at least 2 visual−anchor
        nodes.


### 1. NODE TYPES  grounded in ToM theory

Each node "id" = capital−letter prefix + index (D1, A1, ).
"text" field MUST start with the same ID prefix.

We organize nodes along the ToM processing pipeline (Baron−Cohen's
mindreading system + Wellman's belief−desire framework):

#### LAYER 1  Observable Inputs (what can be directly seen/heard)
These are the RAW EVIDENCE that a mindreader uses to infer
        mental states.

 A  Action
    Intentional physical behaviour: gestures, body language, posture
    shifts, approach/avoidance movements. In ToM theory, actions are
    the PRIMARY OBSERVABLE OUTPUT of the belief−desire
        system  we
    reverse−engineer mental states from watching what people DO.
    Include micro−actions (lip−biting, gaze−aversion, fidgeting) that
    reveal inner states.
(The following is omitted)

#### LAYER 2  Mental States (the core of ToM reasoning)
These are UNOBSERVABLE internal states inferred from Layer 1
        evidence.
The Wellman belief−desire model: Perception  Belief  Desire  Intent
 Action  Emotion (feedback). You MUST annotate richly here.

 B  Belief  (Wellman's epistemic component)
    What a character takes to be true about the world, themselves, or
    others' minds.
    LEVELS OF BELIEF (annotate which level applies):
      1st−order: "X believes P" (e.g. "Craig believes this is a
        normal workplace")
      2nd−order: "X believes that Y believes P" (e.g. "Lotte
        believes Craig thinks she is Maxine")  this is classic ToM.
      FALSE BELIEF: when the character's belief contradicts reality.
        This is the gold standard of ToM tasks. Clearly mark when a
        belief is false and WHY the character holds it despite reality.

\textbf{(The following is omitted)}


### 2. KEYFRAME TIMESTAMP PRECISION  (CRITICAL)

For EVERY node, the `timestamp`  field must point to the EXACT
        moment in
the video where the evidence is strongest. This is used to extract
keyframes for bounding−box annotation.

For visual nodes (A, P, D, E with visual evidence):
 − Pick the SINGLE BEST FRAME where the
        action/expression/object is
  most clearly visible.
 − Example: if a character's anger builds from 10s to 15s, and they
  slam their fist at 14.2s  set timestamp=14.2
 − For facial expressions: pick the PEAK of the expression, not the
  start. Example: "furrowed brow"  the frame where the furrow is
  deepest.

For audio/inference nodes:
 − Set timestamp to when the TRIGGER is visible. Example: for
  "B1: He believes she is lying" (inference from P1+L2), set
  timestamp to the moment he shows suspicion on his face.


### 3. EVIDENCE MODALITY + GROUNDING

Every node MUST have `evidence_modality`: "visual" / "audio" /
        "inference".

For "visual" nodes:
 − `grounding_query`: CONCRETE phrase for a local detection
        model.
  NEVER USE CHARACTER NAMES  use generic descriptors
        instead.
  Good: "man with glasses and messy hair, furrowed brow",
        "woman in white sweater holding cigarette, looking down",
        "person's clenched fist on table", "TV screen showing hunched
        figure"
  Bad: "Craig's expression", "Maxine sitting" (names won't work!)
  Bad:  "confusion", "tension" (too abstract)
 − `is_visual_anchor`: true if this is a clear, unambiguous visual
  moment. Each clip needs 2 visual anchors.
 − `visual_cue`: 3−8 word description of what makes it visually
  identifiable (e.g. "clenched jaw, narrowed eyes").

For "audio" nodes:
 − `audio_grounding`: {"transcript_excerpt": "...", "tone": "..."}
 − You MAY add a weak `grounding_query`  (e.g. speaker's face) with
  `"weak_visual": true`.

For "inference" nodes:
 − `inference_from`: list of node IDs that serve as premises.
 − Explain the reasoning chain: WHY does A1+L2 lead to B1?


### 4. DEEP PSYCHOLOGICAL ANALYSIS  (MANDATORY)

This is a ToM benchmark, NOT a scene description dataset. You
        MUST go
beyond surface−level annotation:

 **Belief nodes (B)**: Don't just say "believes X". Explain the
 epistemic basis: "B1: Craig believes this is a normal office job,
 based on the fact that he accepted the position without questioning
 the unusual floor number (inference from L1, K1)."

...


### 5. CLIP SELECTION RULES

 MINIMUM {clip_min}s, target 20−45s. Short clips miss the full arc.
 Each clip must contain a PSYCHOLOGICAL ARC: a trigger event
        that
 causes mental state changes that lead to behavioral consequences.
 Each clip must have 5 nodes total, 2 mental−state nodes (B/I/R/E),
 and 2 visual−anchor nodes.
 Prefer FEWER, LONGER, RICHER clips over many short shallow
        ones.
 Non−overlapping clips only. Leave gaps if needed.


### 6. OUTPUT FORMAT  (strict JSON)

```json
{
  "source_video_duration": 180.5,
  "clips": [
    {
      "clip_id": 1,
      "segment": {
        "start_time": 12.0,
        "end_time": 38.5,
        "reason_for_selection": "Craig's disorientation arc: trigger (low
            ceiling)  belief shift  visible anxiety"
      },
      "script": "In this clip ",
      "characters": [
        {"id": "c1", "name": "Craig"},
        {"id": "c2", "name": "Don"}
      ],
      "nodes": [
```
```

```
      {
        "id": "P1",
        "type": "percept",
        "text": "P1: Craig notices the abnormally low ceiling as he
         enters, ducking slightly.",
        "character": "c1",
        "timestamp": 12.4,
        "end_time": 13.5,
        "evidence_modality": "visual",
        "is_visual_anchor": true,
        "visual_cue": "man ducking under low ceiling",
        "grounding_query": "man in suit ducking his head under low
         ceiling",
        "audio_grounding": null,
        "inference_from": null,
        "caused_by": null
      },...
    ]
  }
 ]
}
```


C.3 Visual Node Verification

For each visual-modality node from Stage 1, a ± 3 s segment around its annotated timestamp is sent back to Qwen-2.5-VL-72B for verification. Nodes scoring below 5 are downgraded to inference modality.

Prompt for Visual Node Verification

```
You are a strict keyframe verification assistant for a Video
Theory−of−Mind benchmark.

You will receive a short video segment and a list of visual queries.
For each query, watch the segment and answer:

1. visible (true/false):
  Can you clearly see the described content at any point in this
      segment?

2. best_time (float):
  If visible, give the exact second within the segment where the content
  is most clearly visible, such as the peak of an expression or the
  clearest view of an action.

3. score (0−−10):
  10  = perfect match, crystal clear
  7−−9 = clearly visible, minor differences
  5−−6 = partially visible, some ambiguity
  3−−4 = barely visible, mostly guessing
  0−−2 = not visible or contradicts the description

4. reason:
  Provide a one−sentence explanation.

Rules:
− Watch the entire segment.
− For facial expressions, look for the peak frame.
− Be strict: assign score < 5 if you cannot clearly see the content.
− Do not hallucinate.

Output strict JSON:
{"results":[{"query_idx":0,"visible":...,"best_time":...
,"score":...,"reason":"..."}, ...]}
```

C.4 Bounding-Box Annotation and Inspection

For each verified visual node, local Qwen3-VL-8B is used to produce a candidate grounding box from the node's keyframe and grounding query. We then use Qwen2.5-VL-72B model for bounding-box quality inspection and calibration. Coordinates are normalized across three possible scales: normalized [0,1], absolute pixel, and Qwen-VL 1000-scale.

Prompt for Local Bounding-Box Annotation

```
You are a visual grounding assistant.

Given an image of size W x H pixels and a list of queries, output
bounding boxes.

For each query, provide [x1, y1, x2, y2] in absolute pixels,
not normalized.

Coordinate definition:
− x1, y1 = top−left corner
− x2, y2 = bottom−right corner

Boxing rules:
− For faces or expressions, box the face tightly.
− For actions or postures, box the full body.
− If the queried entity is not present, set bbox to null.

Output JSON only:
{"results":[{"idx":0,"found":true,"bbox":[x1,y1,x2,y2]}, ...]}
```

C.5 VLM QA Generation

Prompt for VLM QA Generation

```
You are an expert psychologist and benchmark engineer constructing a
∗∗Video Theory−of−Mind (ToM)∗∗ evaluation dataset.

You will receive:
 (a) A set of ∗∗video frames∗∗ from a short clip (~20−45 s).
 (b) A ∗∗structured annotation∗∗ of the clip: characters, psychological
      nodes
    (with timestamps, evidence modality, bounding−box grounding,
      causal
    links), and causal subchains.
 (c) Optionally, ∗∗legacy QA examples∗∗ from a prior dataset  use
      them as a
    quality reference for writing style and distractor quality, NOT for
    copying answers.

Your task: generate a comprehensive set of ∗∗5−way multiple−choice
      questions∗∗
covering ALL 9 categories defined below.
CARDINAL RULE: VIDEO−DEPENDENCY

Every question you write MUST require watching the video to answer
      correctly.
A text−only language model with NO access to the video should NOT
      be able to
choose the correct answer.

HOW TO ACHIEVE THIS  "Deictic Anchoring":
GOOD: Reference video moments, don't describe them.
    "At around 15 seconds, what does the person in the chair do?"
    "Based on the facial expression visible at the start of the clip, ..."
    "What visual change occurs after the dialogue at ~20s?"
BAD: Narrate the video content IN the question stem.
    "Given Craig's initial weary and apathetic state upon sitting down,
      ..."
    "After seeing employees hunched over and hearing the narrator ask
      ..."
    (These questions give away the answer without needing the video!)

 RULES:
 a. Question stems should POINT TO a moment ("around Xs", "in
      the first
    few seconds", "right after the character speaks") rather than
    DESCRIBING what happens at that moment.
 b. If the question must mention a mental state, keep it MINIMAL
      and ask
    for its CAUSE or CONSEQUENCE, not for the state itself.
```

c. NEVER put the correct answer's key information in the question stem.
d. Distractors must be plausible to someone who has NOT seen the video.
A text–only model should find 2 distractors equally appealing.

QA TAXONOMY 9 Categories in 3 Directions

FORWARD REASONING
Given earlier events OBSERVED IN THE VIDEO, predict what follows. The question should POINT to a moment, not describe the full context.

F1 ACTION PREDICTION
"After what happens at ~Xs, what does [character] do?"
The premise is a VIDEO MOMENT, not a narrated state.
The viewer must watch to know what happened at ~Xs.

F2 EMOTION PREDICTION
"How does [character] react emotionally to the event that occurs around ~Xs?"
Do NOT name the event force the viewer to see it.
Distractors: plausible emotions for similar situations.

F3 DIALOGUE PREDICTION
"Given what [character] sees/hears at ~Xs, what would they most likely say next?"
Distractors: things someone COULD say in a similar scene.

\textbf{(The following is omitted)}

1. 5 options per question. Exactly one correct. Mark `answer\_index` (0–based). The correct answer should NOT always be in the same position;
randomize it across questions.

2. **Distractors must be HARD and equally attractive without video.**
At least 2 distractors must be plausible to someone who has NOT seen
the video (they should sound like reasonable answers given only the question text).
At least 1 distractor should be a common–sense alternative the answer you'd guess if you only read the question stem.
NEVER use obviously absurd or irrelevant options.
All 5 options must have SIMILAR length and grammatical structure.
The correct answer must NOT be the longest or most detailed option.
NEVER put qualifiers like "exactly", "definitely" only on distractors while the correct answer uses hedging like "likely" this is a tell–tale pattern.
(The following is omitted)

Output strict JSON:
{"qas":[{...}, ...]}

D QA Design

To evaluate whether models can reason over mental-state trajectories rather than answer isolated recognition questions, we design three families of multiple-choice questions. Each family probes a different direction of the evidence–mind–behavior chain. Forward questions ask models to predict what will happen next from an observed event or mental-state cue. Backward questions ask models to explain why an observed action, belief, or emotional shift occurred by identifying its preceding evidence. Grounded perception questions ask models to localize the visual, temporal, or perspective-specific evidence that supports a ToM inference. Together, these categories test whether a model can move from observable cues to latent mental states and back to behavioral consequences.

Cat.	Direction	Reasoning Target
F1	Forward	Predict the next action after an observed event.
F2	Forward	Predict the character's emotional reaction.
F3	Forward	Predict the next utterance from heard context.
B1	Backward	Explain the cause of a visible behavior.
B2	Backward	Identify evidence that supports a belief.
B3	Backward	Explain the trigger of an emotional change.
G1	Grounded	Identify the visual detail supporting a mental state.
G2	Grounded	Locate the time span of a behavioral or mental-state shift.
G3	Grounded	Identify perspective or information asymmetry.

Table 5: Reasoning targets of the nine QA categories.

Each QA pair is generated from the structured ToM graph and is linked to source nodes, timestamps, and bounding boxes when visual evidence is available. The generator produces five answer options with exactly one correct choice and plausible distractors, so that answering requires grounding the question in the video rather than relying on social priors alone. Tables 5, Fig 5 and Fig 6 summarize the design of the nine QA categories. The taxonomy is organized by reasoning direction rather than surface question form: forward questions predict behavioral consequences, backward questions recover causal or psychological antecedents, and grounded questions verify the visual, temporal, or perspective-specific evidence that supports a ToM inference.

E Quality Control and Validity of the Pipeline

This section provides additional details on the automatic filtering and human verification procedures used to improve the validity of ToM-Video . The goal is to remove samples whose visual evidence, source-node chain, or QA label is ambiguous, weakly grounded, or inconsistent with the video.

E.1 Automatic Quality Filtering

Before human review, we apply automatic checks to three components of the pipeline: ToM nodes,

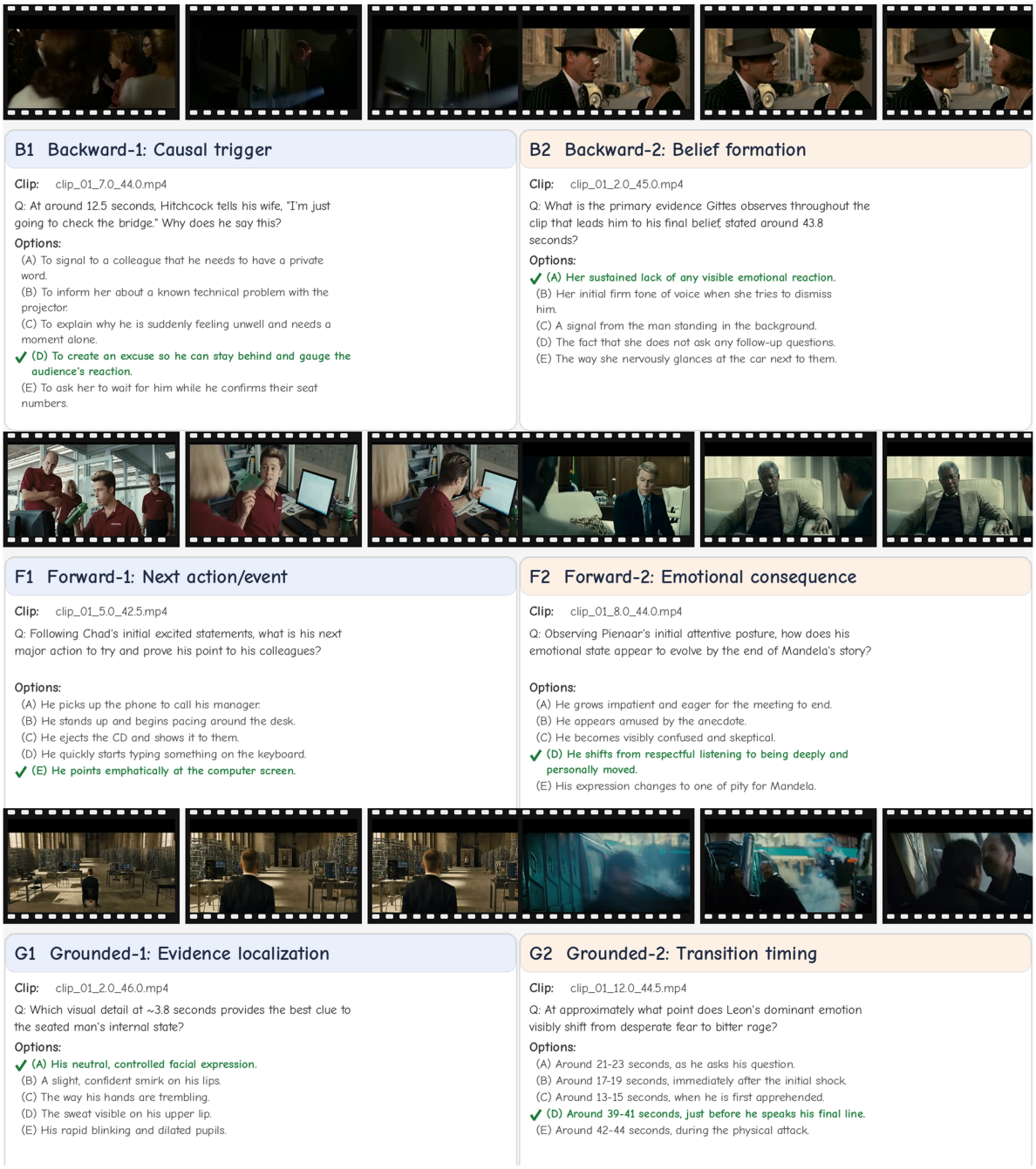

Figure 5: QA Examples

bounding boxes, and QA pairs. For visual nodes, we remove or demote claims that are not clearly visible at the annotated timestamp. A visual node is retained only if the verification model confirms that the described cue is perceivable in a short temporal neighborhood around the predicted timestamp. Nodes that are plausible but not directly visible are converted to inference nodes, while unsupported or contradictory nodes are discarded.

For spatial grounding, we discard bounding boxes that are invalid, degenerate, outside the image canvas, or too small to provide meaningful evidence. This removes grounding artifacts and prevents downstream QA pairs from citing unreliable visual regions.

For QA pairs, we filter samples that are structurally invalid, answerable from text alone, ambiguous, missing source-node chains, or inconsistent with the grounded rationale. We also remove questions that require external plot knowledge rather than evidence available in the clip. These checks are designed to preserve video-dependent ToM reasoning while reducing annotation noise.

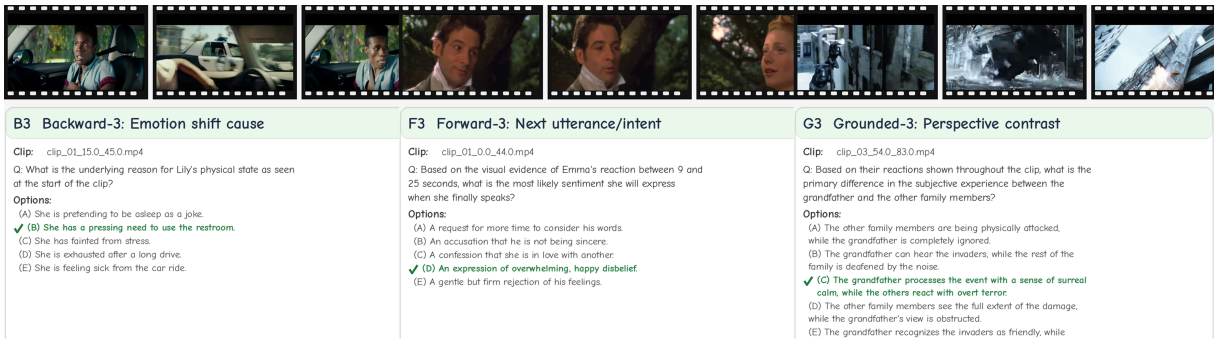

Figure 6: QA Examples (add)

E.2 Human Audit Protocol

Automatic filtering cannot fully resolve subtle social, perceptual, and pragmatic ambiguities in ToM reasoning. We therefore conduct an additional human audit. Ten trained reviewers with backgrounds in cognitive science, linguistics, and AI participate in the verification process. Each candidate sample is independently reviewed by two annotators.

Annotators are shown the video clip, ASR transcript, structured ToM graph, source-node chain, question, answer options, correct answer, grounded rationale, and rendered keyframe with bounding boxes. They evaluate four aspects:

1. **Video and clip quality**: whether the clip is understandable and contains sufficient visual or auditory evidence.

2. **ToM-node correctness**: whether each node correctly describes an observable cue, latent mental state, or social-cognitive relation.

3. **Grounding consistency**: whether each visual node corresponds to a perceivable cue at the cited timestamp and whether the bounding box covers the intended character, object, or region.

4. **QA validity**: whether the question is answerable from the video and ASR without external knowledge, has exactly one correct answer, and is supported by the cited source-node chain.

Items approved by both annotators are retained. Items rejected by both annotators are removed. Disagreements are resolved by a senior adjudicator.

Table 6 reports the cumulative effect of automatic filtering and human verification. Automatic filtering removes hallucinated visual claims, invalid spatial groundings, structurally invalid QA pairs, and weak source-node chains. Human verification further removes items that are visually ambiguous, weakly grounded, or socially under-determined.

Component	Raw	After auto.	After human
Mental-state nodes	87,879	63,412	57,707
L1: Observable	38,642	27,881	25,998
L2: Mental	45,721	32,710	29,375
L3: Social	3,516	2,821	2,334
Multiple-choice QAs	55,016	43,591	41,206

Table 6: Cumulative effect of automatic filtering and human verification on ToM-Video .

E.3 Independent Annotation Validity Study

The construction-time audit described above verifies candidate samples before they enter the final dataset. However, filtering counts alone do not estimate the residual reliability of the retained benchmark. This is especially important for ToM-Video , since the pipeline uses MLLMs for ToM-dense segment selection, node extraction, visual verification, bounding-box grounding, and QA generation. We therefore conduct an independent validity audit on a stratified sample from the final benchmark.

Sampling. We sample $N = 1{,}000$ QA pairs from the final benchmark. The sample is stratified by (i) the nine QA categories, (ii) difficulty, (iii) the dominant evidence modality in the source chain (visual, audio, inference), and (iv) video category. This design ensures that the audit covers visually grounded, dialogue-driven, and inference-heavy cases.

Annotators and blinding. The audit is conducted by two annotators who did not participate in data generation, automatic filtering, or construction-time human review. Annotators are shown the video clip, ASR transcript, question,

answer options, the released answer key, source-node chain, and rendered keyframes with bounding boxes. They are not shown the generation prompt, model identity, or whether an item was corrected during the pipeline. Each item is independently reviewed by both annotators; disagreements are resolved only after agreement statistics are computed.

Audit dimensions. Annotators evaluate four core validity dimensions:

1. **Answer-key validity**: whether the marked answer is uniquely correct and all distractors are clearly less correct.

2. **Source-chain sufficiency**: whether the cited source nodes provide enough evidence to support the answer.

3. **Grounding correctness**: for visual nodes, whether the cited timestamp contains the relevant cue and whether the bounding box covers the intended person, object, or region; for audio nodes, whether the ASR span supports the node.

4. **Mental-state inference validity**: whether the transition from observable evidence to the annotated belief, intention, desire, emotion, knowledge, or non-literal meaning is justified without over-interpretation or external plot knowledge.

Metrics. We report inter-annotator agreement using Gwet's AC1 for binary validity decisions and Krippendorff's α for multi-class error labels. We also report answer-key validity, source-chain F1, timestamp median absolute error (MedAE), Time@3s-human, human-vs-released box IoU, Box@0.5-human, text-only answerability, ASR-only answerability, external-knowledge dependence, and overall residual error rate with 95% confidence intervals.

For source-chain F1, annotators independently select the source nodes they consider necessary for supporting the answer, and we compute F1 between the released source chain and the human-selected chain. For timestamp evaluation, annotators mark the best supporting moment when the released timestamp is judged imprecise. For box evaluation, annotators either accept the released box or draw a corrected box on the rendered keyframe.

We define an item as having a residual error if any of the following holds: the answer key is not uniquely correct, the source chain is insufficient, a required visual or audio grounding is incorrect, the mental-state inference is unjustified, the question is answerable without video, or the answer requires external plot knowledge.

Overall, the independent audit provides a direct estimate of retained-data reliability beyond the cumulative filtering counts. The remaining errors are concentrated in timestamp imprecision, source-chain insufficiency, and mental-state over-inference, while answer-key ambiguity and external-knowledge dependence remain rare. These results suggest that the final benchmark is reliable for evaluating annotation-aligned evidence-grounded ToM reasoning, while also identifying the annotation dimensions that require caution.

E.4 Filtering Statistics

After all filtering stages, ToM-Video contains 57,707 ToM nodes and 41,206 multiple-choice QA pairs.

F Statistical Analysis of ToM-Video

We provide the full statistics underlying the headline numbers in §5. Tables 8 and 9 summarise the corpus and per-category QA counts, while Figures 7–8 visualise ontology coverage and surface statistics.

Distribution of the 57,707 annotated mental-state nodes across the 3-layer / 10-category ontology of §3. The inner ring shows the three theoretical layers (*Observable Inputs*, *Mental States*, *Social-Cognitive Processes*); the outer ring expands into the ten node types. Layer 2 (Mental) accounts for the majority of supervisory signal (50.9%), confirming that ToM-Video is dominated by mental-state inference rather than purely perceptual content. (a) Source-video duration is concentrated around two minutes (mean 122.6 s), well above standard short-form VideoQA datasets. (b) Each video contributes on average 3.11 psychologically interactive clips, ensuring multi-scene coverage per source. (c) Questions average 19.95 words — longer than MovieQA, TVQA and SocialIQ — reflecting the deictic anchors mandated by the cardinal video-dependency rule. (d) Distractor and correct-answer lengths are nearly identical (7.80 vs. 8.46 words), reducing the well-known length-shortcut bias.

Audit dimension	Agreement	Valid / Score
Answer key uniquely correct	AC1 = 0.91	98.6% [97.7, 99.2]
Source chain sufficient	AC1 = 0.84	96.8% [95.5, 97.7]
Mental-state inference justified	AC1 = 0.82	96.2% [94.8, 97.2]
Visual timestamp correctness	α = 0.81	MedAE = 0.8s; Time@3s = 94.7%
Box correctness	α = 0.79	IoU = 0.81; Box@0.5 = 93.8%
Source-chain agreement	–	F1 = 0.91
Question-only answerable	AC1 = 0.86	2.4% [1.6, 3.5]
ASR-only answerable	AC1 = 0.82	5.8% [4.5, 7.4]
External plot knowledge required	AC1 = 0.88	1.2% [0.7, 2.1]
Overall residual error	–	5.9% [4.6, 7.5]

Table 7: Independent validity audit on a stratified sample of final ToM-Video items. Numbers are reported as percentages unless otherwise noted. CI denotes a 95% confidence interval. Overall residual error is computed as the item-level union of answer-key invalidity, source-chain insufficiency, unjustified mental-state inference, incorrect required grounding, question-only answerability, ASR-only answerability, and external-knowledge dependence.

Property	Value
Source videos	1,487
Video duration, avg./median	122.6 s / 128.0 s
Psychologically interactive clips	4,666
Clips per video, avg.	3.14
Annotated mental-state nodes	57,707
Nodes by layer, L1/L2/L3	25,998 / 29,375 / 2,334
Multiple-choice QA pairs	41,206
QAs per video, avg.	27.7
Question length, avg.	19.95 words
Answer length, corr./dist.	8.46 / 7.80 words

Table 8: Top-level corpus statistics of ToM-Video.

Note. L1 = Observable, L2 = Mental, and L3 = Social.

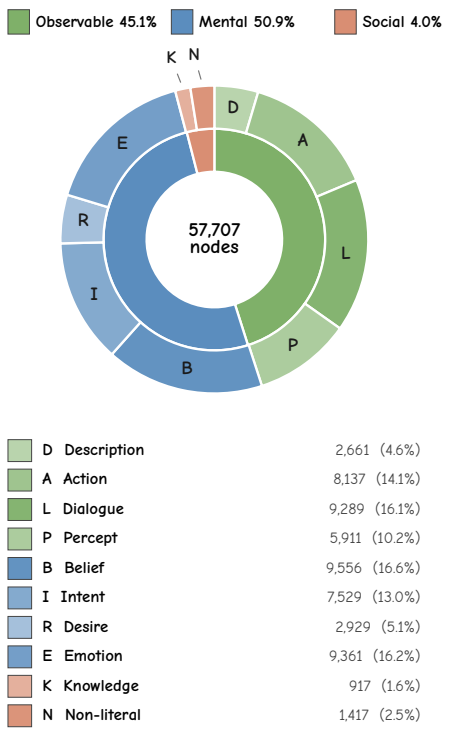

Figure 7: Ontology

Take-aways. (i) Layer 2 mental-state nodes (Belief, Intent, Desire, Emotion) constitute over half of all annotated nodes (50.9%), validating that ToM-Video is dominated by mental-state inference rather than passive scene description. (ii) Question lengths (mean 19.95 words) are roughly 1.5× those of standard VideoQA datasets, consistent with the deictic anchoring required by the cardinal video-dependency rule.

G Training configuration

We adopt a three-phase curriculum: SFT → RL-Grounding → RL-ToM, with each phase initialized from the previous checkpoint. All phases share the Qwen3-VL-8B backbone, AdamW optimizer with cosine schedule, ZeRO-3 sharding without CPU offload, FlashAttention-2, bf16 mixed precision, and run on 4×H800 (80 GB) GPUs.

GPU Hours. We report the computational budget in H800 GPU-hours. All three stages are trained on 4×H800 (80 GB) GPUs. The SFT stage takes approximately 10 wall-clock hours, corresponding to 40 H800 GPU-hours. RL-Grounding is trained until the reward curve saturates, taking approximately 2 days (48 wall-clock hours), corresponding to 192 H800 GPU-hours. RL-ToM takes approximately 1.5 days (36 wall-clock hours), corresponding to 144 H800 GPU-hours. In total, the full INSPIRE

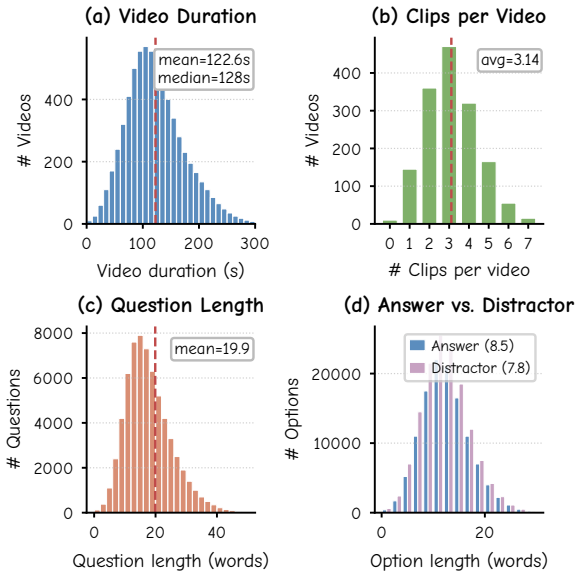

Figure 8: Surface statistics of ToM-Video.

Category	Total
F1 — Action prediction	4,663
F2 — Emotion prediction	4,657
F3 — Dialogue prediction	3,914
B1 — Causal explanation	4,688
B2 — Belief attribution	4,636
B3 — Emotion attribution	4,659
G1 — Visual evidence	4,673
G2 — Temporal localization	4,659
G3 — Perspective taking	4,657
Total	**41,206**

Table 9: Per-category QA counts.

training pipeline requires approximately 376 H800 GPU-hours. These numbers include forward/backward training and on-policy generation during RL, but do not include data construction, annotation, or baseline evaluation.

Training data allocation. We allocate data according to the objective of each training phase. To estimate shortcut difficulty, we run a blind text-only test with GPT-4o five times per training question, using only the question and options. A sample is labeled easy if GPT-4o answers it correctly at least once, and hard otherwise. SFT uses all 36,893 training examples to learn the output schema and the overall task distribution. RL-Grounding then uses 14,984 grounding-heavy hard examples with explicit timestamp and box supervision, focusing on local evidence localization. RL-ToM uses 13,464 hard examples requiring source-chain-supported mental-state inference, aligning grounded traces with ToM-oriented answer selection. This staged allocation separates format learning, evidence localization, and mental-state decision-making.

RL-Grounding. Starting from the SFT checkpoint, we apply GRPO (Shao et al., 2024) with the structure and grounding components of the INSPIRE reward, using weights $(\mu_{\text{str}}, \mu_{\text{grd}}) = (1.0, 1.5)$. The temporal matching window of R_{grd} follows the curriculum schedule $\Delta(\gamma) = \max\big(1.0,\ 4.0\,(1-\gamma)\big)$ s defined in Eq. (2), where $\gamma \in [0, 1]$ denotes normalized training progress: the window starts at 4 s and gradually tightens to 1 s by the end of training, so that early approximate timestamps still receive partial credit while the policy is progressively pushed toward precise temporal grounding. We sample groups of $G = 3$ completions per prompt with $T = 1.0$, top-$p = 0.95$. The KL coefficient is $\beta = 0.01$ and asymmetric clipping uses $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}}) = (0.20, 0.28)$.

RL-ToM. Initialized from the RL-Grounding checkpoint, RL-ToM swaps the active reward mixture to grounding, content, and consistency, with weights $(\mu_{\text{grd}}, \mu_{\text{cnt}}, \mu_{\text{cns}}) = (0.3, 1.0, 1.0)$, so that the model is rewarded for ToM-aware answer correctness and source-trace consistency while the smaller grounding weight stabilises spatial faithfulness. The group size is enlarged to $G = 8$ and the KL coefficient to $\beta = 0.04$ for richer advantage estimates and stronger regularisation toward the RL-Grounding policy. All other RL settings (LR 1×10^{-6}, 0.05 warm-up, cosine, effective batch 8, prompt 4,096 / completion 768 tokens, ∼401K pixels/frame, two epochs, GSPO importance weighting and asymmetric clipping $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}})$ as in RL-Grounding) follow RL-Grounding. For R_{cns}, we use the fixed temporal tolerance $\delta = 4$ s specified in Eq. (5); R_{grd} continues to follow the curriculum schedule $\Delta(\gamma)$ from RL-Grounding.

H Reward Implementation Details

Sampling and decoding. At every optimizer step the policy generates G continuations per prompt with $T = 1.0$, top-$p = 0.95$, and at most max_completion_length new tokens. Generation runs with gradient checkpointing temporarily disabled to preserve the KV cache of Qwen3-VL during multi-step decoding.

Group-relative advantage and reward weighting. For each prompt we sample a group of G completions and evaluate the active reward components separately. Each component is multiplied

Hyper-parameter	SFT	RL-Grd	RL-ToM
Base ckpt.	Qwen3-VL-8B	SFT	RL-Grd
Optimiser	AdamW	AdamW	AdamW
LR	$1e{-}5$	$1e{-}6$	$1e{-}6$
Schedule	cosine	cosine	cosine
Warm-up ratio	0.03	0.05	0.05
Weight decay	0.01	0.01	0.01
Grad-norm clip	5.0	5.0	5.0
Epochs	1	2	2
Eff. batch	$1\times16\times4{=}64$	$1\times2\times4{=}8$	$1\times2\times4{=}8$
Prompt tok.	8192	4096 (L-trunc.)	4096 (L-trunc.)
Completion tok.	—	768	768
Pixels / frame	$\sim$200K	$\sim$401K	$\sim$401K
Sampling	—	$T{=}1,\ p{=}0.95$	$T{=}1,\ p{=}0.95$
Group size G	—	3	8
KL β	—	0.01	0.04
Clip $(\varepsilon_{\text{lo}}, \varepsilon_{\text{hi}})$	—	(0.20, 0.28)	(0.20, 0.28)
Importance weight	—	GSPO (seq-mean)	GSPO (seq-mean)
Active rewards	CE	$R_{\text{str}}, R_{\text{grd}}$	$R_{\text{grd}}, R_{\text{cnt}}, R_{\text{cns}}$
Reward weights μ	—	1.0, 1.5	0.3, 1.0, 1.0
Hardware	4×H800	4×H800	4×H800
DeepSpeed	ZeRO-3	ZeRO-3	ZeRO-3
Attention	FlashAttention-2	FlashAttention-2	FlashAttention-2
Precision	bf16	bf16	bf16

Table 10: Training hyper-parameters per stage. "Eff. batch" denotes per_device × grad_accum × 4 GPUs. GSPO denotes sequence-level importance weighting; clip is the asymmetric DAPO clip; KL uses the Schulman k_3 estimator with the log-ratio clamped to $[-10, 10]$.

by its scalar weight μ_* *inside* the reward function, and the trainer then sums the weighted components into a scalar reward r_i per completion. We follow Dr. GRPO (Liu et al., 2025) and take the group-mean–centered advantage

$$A_i \;=\; r_i \;-\; \tfrac{1}{G}\sum_{j=1}^{G} r_j, \tag{10}$$

without dividing by the group standard deviation. Pre-multiplying each reward by μ_* rather than rescaling the sum after normalization lets the relative magnitude of fast-saturating signals (e.g. format compliance) be controlled at the source while still benefiting from the variance-reduction of the group baseline.

GRPO loss. The vanilla group-relative policy optimization objective (Shao et al., 2024) is

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) \;=\; & \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \\ & \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\big(r_{i,t}(\theta)\,\hat{A}_{i,t}, \\ & \quad \text{clip}(r_{i,t}(\theta), 1{-}\varepsilon, 1{+}\varepsilon)\,\hat{A}_{i,t}\big) \\ & \qquad - \;\beta\,\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\Big)\Bigg],\end{aligned} \tag{11}$$

with token-level importance ratio and group-normalised advantage

$$r_{i,t}(\theta) \;=\; \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \tag{12}$$

$$\hat{A}_{i,t} \;=\; \frac{R_i \;-\; \text{mean}\big(\{R_j\}_{j=1}^{G}\big)}{\text{std}\big(\{R_j\}_{j=1}^{G}\big)}. \tag{13}$$

Our implementation differs from Eq. (11) in four respects.

(i) Sequence-level importance ratio (GSPO (Zheng et al., 2025)). With completion mask $m_t \in \{0, 1\}$ and per-token log-ratio $\delta_t = \log \pi_\theta(o_t) - \log \pi_{\theta_{\text{old}}}(o_t)$, the token-level ratio $r_{i,t}(\theta)$ in Eq. (11) is replaced by the per-response ratio

$$s_i(\theta) \;=\; \exp\Big(\tfrac{\sum_t m_t\,\delta_t}{\max(1, \sum_t m_t)}\Big), \tag{14}$$

broadcast to every token of completion i.

(ii) Asymmetric DAPO clip (Yu et al., 2025). The symmetric threshold ε is replaced by the pair $(\varepsilon_{\text{low}}, \varepsilon_{\text{high}}) = (0.20, 0.28)$, giving the clip $\text{clip}\big(s_i,\ 1{-}\varepsilon_{\text{low}},\ 1{+}\varepsilon_{\text{high}}\big)$. The wider upper bound favours upward exploration while the tighter lower bound contains destructive negative updates.

(iii) Dr. GRPO advantage (Liu et al., 2025). As detailed in the previous paragraph, $\hat{A}_{i,t}$ is replaced by the mean-centred group advantage

$$A_i \;=\; R_i \;-\; \tfrac{1}{G}\sum_{j=1}^{G} R_j, \tag{15}$$

without variance normalisation, and broadcast to every token of completion i.

(iv) KL via Schulman k_3 with clamping. The KL penalty is estimated on-policy by the unbiased k_3 estimator

$$\begin{aligned}\widehat{\text{KL}}_t \;&=\; e^{x_t} \;-\; x_t \;-\; 1, \\ x_t \;&=\; \text{clamp}(\log \pi_{\text{ref}} - \log \pi_\theta,\ -10,\ 10)\,.\end{aligned} \tag{16}$$

The reference model π_{ref} is frozen at the SFT (Phase 2) or RL-Grounding (Phase 3) checkpoint and prepared under DeepSpeed ZeRO-3.

Combining (i)–(iv), and noting that s_i and A_i are constant across the tokens of completion i, the implemented loss takes the per-response form

$$\begin{aligned}\mathcal{L}(\theta) \;=\; \frac{1}{B}\sum_{i=1}^{B}\Big\{ & -\min\big(s_i\,A_i,\ \text{clip}(s_i, \\ & 1{-}\varepsilon_{\text{low}}, 1{+}\varepsilon_{\text{high}} A_i\big) \;+\; \beta\,\widehat{\text{KL}}^{(i)}\Big\},\end{aligned} \tag{17}$$

where the per-response KL average is

$$\overline{\widehat{\mathrm{KL}}}^{(i)} = \frac{\sum_t m_t^{(i)} \widehat{\mathrm{KL}}_t^{(i)}}{\max\big(1, \sum_t m_t^{(i)}\big)}. \tag{18}$$

The outer batch mean together with the per-response length normalisation of the KL term recovers GRPO-style aggregation; we do not adopt the full DAPO global-token-mean loss, since it under-weights short completions, which dominate the INSPIRE F/B categories.

Source-node fidelity. Let $\mathcal{S}_q^\star$ be the gold source-node set for question q, and let $\widehat{\mathcal{S}}(\tau)$ be the set of node identifiers extracted from the reasoning trace τ. We define

$$\rho_{\mathrm{src}}(\tau, q) = \begin{cases} \mathrm{F}_1\Big(\widehat{\mathcal{S}}(\tau), \mathcal{S}_q^\star\Big), & \mathcal{S}_q^\star \neq \emptyset, \\ \text{N/A}, & \mathcal{S}_q^\star = \emptyset. \end{cases} \tag{19}$$

The predicted set $\widehat{\mathcal{S}}(\tau)$ is extracted with a deterministic regular-expression matcher (`\b([A-Z]\d+)\b`) over explicit node identifiers such as B1, I2, and L1. Duplicate references are collapsed before computing F_1, and unmatched predicted nodes are counted as false positives.

Perspective detection. The perspective score is $\rho_{\mathrm{persp}}(\tau) = \mathbb{1}[\tau \text{ matches a fixed perspective pattern}]$. The pattern set covers expressions of belief, knowledge, awareness, perception, and asymmetric information access, including *believes*, *thinks*, *knows*, *sees*, *notices*, *is aware/unaware*, *from X's perspective*, and *without X knowing/realizing*.

ρ_{persp} is intentionally a deterministic surface-form proxy rather than a learned classifier, to keep the reward non-adversarial and the optimisation signal interpretable. Two design choices prevent the policy from gaming this proxy via keyword spamming. First, ρ_{persp} enters the reward only through R_{cnt} (Eq. (4)), which is gated by final-answer correctness ($\mathbb{1}[o{=}o^\star]$) and weighted at $\lambda_p{=}0.2$: a trace that emits perspective keywords without selecting the correct option receives zero reward, and even with the correct answer the perspective term contributes at most $0.2 \,/\, 1.0$ of R_{cnt}. Second, we cross-validate the surface proxy against human chain validity $\bar{\chi}$ (Table 3, Appendix K.4), which explicitly rejects traces that cite no grounded evidence or attribute mental states without verifiable cues. After RL-ToM, $\bar{\chi}$ rises from 60.3 to 73.6 on SocialIQ and from 59.4 to 81.2 on TOM-VIDEO-VAL. Because $\bar{\chi}$ is annotated under double-blind protocol on randomly ordered traces with formatting stripped, it cannot be inflated by keyword surface form alone; the joint improvement of ρ_{persp} and $\bar{\chi}$ therefore indicates that the policy learns substantive perspective evidence rather than the surface pattern itself.

I Data leakage check

We adopt a strict *movie-level* train/val split: the 1,339 source videos used for SFT and the 148 source videos used for evaluation are disjoint, with **0% overlap** in (i) source video ID, (ii) extracted clip files, (iii) question text, and (iv) (clip, question) pairs. At the same time, we also carefully checked the possibility of overlap between external datasets and the video training set in this paper. videoMME has no overlapping videos, and SocialIQ has removed a few overlapping videos. Other external social scene datasets, such as MomentS, have more overlap and are not included in the evaluation.

J Grounded Process Evaluation

J.1 Process Studies

To complement answer accuracy with an evidence-faithfulness view, we evaluate the intermediate grounding process on TOM-VIDEO-VAL ($N{=}2{,}162$). Each response is parsed into a predicted timestamp $\hat{t}_i$, box $\hat{b}_i$, and source-node set $\hat{\mathcal{S}}_i$; no gold rationales are given at test time. Let $t_j^\star, b_j^\star, \mathcal{S}_i^\star$ denote their gold counterparts and $o_i, o_i^\star$ the predicted/gold answer letters. We report five metrics, all averaged over N (invalid responses contribute 0):

$$\begin{aligned}
\text{Trace} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{1}\Big[\hat{\mathcal{U}}_i \text{ valid}\Big], \\
\text{Time@3s} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{1}\Big[\min_j |\hat{t}_i - t_j^\star| < 3\,\mathrm{s}\Big], \\
\text{Box@0.5} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{1}\Big[\mathrm{IoU}(\hat{b}_i, b_{j^\star}^\star) > 0.5\Big], \\
\text{Src-F1} &= \tfrac{1}{N}\textstyle\sum_i F_1\Big(\hat{\mathcal{S}}_i, \mathcal{S}_i^\star\Big), \\
\text{GAcc.} &= \tfrac{1}{N}\textstyle\sum_i \mathbb{1}\Big[\hat{\mathcal{U}}_i \text{ valid} \wedge o_i{=}o_i^\star \wedge \mathrm{Match}_i\Big].
\end{aligned}$$

A grounding unit $\hat{\mathcal{U}}_i$ is valid iff it contains a parseable $\hat{t}_i$ (plus $\hat{\mathrm{obj}}_i, \hat{b}_i$ when visual evidence is required); Match_i holds when the cited source chain matches $\mathcal{S}_i^\star$.

The gap between answer accuracy (76.2%) and GAcc. (12.53%) on RL-ToM shows that picking

Model	Trace	Time@3s	Box@0.5	Src-F1	GAcc.
Base	4.67	2.31	0.00	0.00	0.00
+SFT	57.96	55.12	42.48	24.33	7.86
+RL-Grd	86.22	58.99	50.68	34.15	11.75
+RL-ToM	**87.65**	**63.52**	**52.75**	**35.41**	**12.53**

Table 11: Process-level evaluation on ToM-Video-val. Source-node credit and GAcc. are counted only on responses with a valid structured trace.

the right option does not imply a faithful evidence chain. SFT lifts Trace from 4.67 to 57.96 (format learning); RL-Grounding nearly saturates it (86.22) while adding +3.9/+8.2 to Time@3s/Box@0.5; RL-ToM tops all five metrics.

J.2 Ablation Studies

Table 12 ablates the visual input at inference time on RL-ToM. Permuting frame order ("Shuffled") causes the largest drop on G2 ($50.3 \rightarrow 28.6$), confirming that INSPIRE relies on the temporal ordering of frames rather than per-frame statistics, with forward-inference categories (F1–F3) also dropping sharply. Removing the visual stream entirely ("ASR only") cuts overall accuracy by 18.5 points and is steepest on G1–G3, showing that the model's gains cannot be explained by language-only shortcuts. Doubling the frame budget from 32 to 64 yields only marginal overall gain (+1.1, mostly from +4.7 on G2), indicating that 32 frames are already close to the input-side ceiling on ToM-Video clips.

K Additional External Generalization Results

We provide supplementary external evaluation results to examine whether the reasoning behavior learned from ToM-Video transfers beyond the in-domain benchmark. We evaluate Qwen3-VL-8B and Qwen3-RL-ToM on two external settings: the short-video subset of Video-MME (Fu et al., 2025) and the accessible subset of Social-IQ 2.0 (Wilf et al., 2023). These evaluations are not used for training or model selection.

K.1 Evaluation Setting

For both external benchmarks, we evaluate the models as multiple-choice video question-answering systems. Each input consists of the video, the question, and the corresponding answer options. The final score is computed only from the predicted multiple-choice answer. No annotated keyframes, bounding boxes, source nodes, or reasoning traces are provided at test time, and intermediate reasoning text is not used for scoring.

Generation is performed with deterministic greedy decoding:

```
max_new_tokens = 512
do_sample = False
use_cache = True
```

The model is loaded with the following inference configuration:

```
torch_dtype = torch.bfloat16
device_map = "auto"
attn_implementation = "flash_attention_2"
```

For scoring, the generated output is post-processed, which extracts the predicted option letter from the model response. A prediction is counted as correct only when the extracted option letter exactly matches the ground-truth option label. The reported metric is accuracy:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = y_i\right], \tag{20}$$

where $\hat{y}_i$ denotes the extracted predicted option letter for example i, y_i denotes the corresponding ground-truth option label, and N is the number of evaluated examples. No partial credit is assigned.

K.2 Video-MME

We first evaluate on the short-video subset of Video-MME (Fu et al., 2025), defined as videos with duration no longer than 60 seconds. Since the current evaluation batch mainly contains short clips, the reported results should be interpreted as accuracy on the Video-MME short-video subset rather than the full Video-MME benchmark score.

As shown in Table 13, Qwen3-RL-ToM improves over the Qwen3-VL-8B backbone, increasing overall accuracy from 73.16% to 75.76%. The gain is especially large on Action Reasoning, where accuracy increases from 54.55% to 72.73%. This suggests that the grounding-aware ToM training transfers most clearly to questions requiring event tracking and temporal evidence integration.

K.3 Social-IQ 2.0

We additionally evaluate on the accessible subset of Social-IQ 2.0 (Wilf et al., 2023). Social-IQ 2.0 focuses on multimodal social interaction understanding.

Setting	B1	B2	B3	F1	F2	F3	G1	G2	G3	Acc	Δ
Full (32 frames, default)	86.0	82.6	83.0	77.3	84.2	80.1	87.4	50.3	88.1	**76.2**	–
Shuffled frames (32)	80.7	76.6	78.6	70.2	74.7	68.4	82.0	28.6	81.7	66.0	−10.2
ASR only	74.9	71.1	69.9	59.9	62.1	53.9	72.1	25.1	72.9	57.7	−18.5
64 frames	86.0	82.6	83.0	77.3	84.2	80.9	87.4	55.0	88.5	77.3	+1.1

Table 12: Video-input ablation on TOM-VIDEO-VAL. Backbone: Qwen3-VL-8B + INSPIRE (RL-ToM). "Shuffled" randomly permutes the order of the 32 sampled frames at inference; "ASR only" removes the visual stream and keeps the ASR transcript.

Model	Video-MME Short	
	Overall	Act. Reason.
Qwen3-VL-8B (Base)	73.16	54.55
Qwen3-RL-ToM (Ours)	**75.76**	**72.73**
Δ	+2.60	+18.18

Table 13: Supplementary external results on the short-video subset of Video-MME. We report multiple-choice accuracy (%).

Model	Social-IQ 2.0	
	Overall	Why-Does
Qwen3-VL-8B (Base)	63.41	61.31
Qwen3-RL-ToM (Ours)	**65.12**	**64.96**
Δ	+1.71	+3.65

Table 14: Supplementary external results on the accessible subset of Social-IQ 2.0. We report multiple-choice accuracy (%).

Table 14 shows that Qwen3-RL-ToM also improves on Social-IQ 2.0, increasing overall accuracy from 63.41% to 65.12% and Why-Does accuracy from 61.31% to 64.96%. These results provide additional evidence that TOM-VIDEO training improves social reasoning beyond the in-domain benchmark, especially for questions that require explaining why a person behaves in a certain way.

K.4 Human Verification Protocol for Chain Validity

We draw n=200 instances from the shared-correct pool by stratified random sampling over ToM sub-skills (belief, desire, intention, emotion, temporal / evidence localization), giving a margin of error of about ±7% at 95% confidence on a binary proportion in the observed range. Two annotators (graduate students with prior experience on multimodal video QA) labeled each instance independently after a 30-item pilot calibration: they see the clip, the question with the gold option, and the two reasoning traces with model identity and trace order randomized and surface formatting stripped, so the source is not recoverable from text style. Each trace receives three binary labels: *evidence validity* (every concrete claim—timestamp, box, action, gaze, quoted speech—is verifiable in the clip), *chain validity* (it cites at least one grounded evidence item, attributes a mental state to a specific agent, links that mental state to the cited evidence, and reaches the answer through this inference rather than from option text alone; length carries no positive weight), and *hallucinated evidence* (the trace contains at least one fabricated, video-checkable claim). Inter-annotator agreement is Cohen's κ=0.71 for chain validity (0.74 for evidence validity, and 0.79 for hallucinated evidence.); the 31% of items with any disagreement were adjudicated by a third annotator under the same blinded protocol, and $\bar{\chi}$ in Table 3 is computed from these adjudicated labels.

L INSPIRE Result Case Study

L.1 Baseline VS SFT VS RL-ToM

Figure 9 traces the three-stage curriculum on a single hard G2 example. **SFT** emits a syntactically valid trace, but its predicted box [0,134,560,211] is a thin strip above Roy's head—the model has learned the trace *schema* but not its visual semantics—and it anchors the transition on the *first* downward glance at 10.5 s, leading to (D) 10–13 s. **RL-Grounding** moves the box onto Roy's face and cites a later cue at 16.5 s, but still equates "Roy looking down" with "transition complete", stopping one beat short of the silent phase and selecting (A) 14–16 s. **RL-ToM** keeps a tight face box and extends the chain to E2: at 18.5–20.0 s Roy is somber and no longer speaking, which is the actual end of verbal deflection; the predicted window 18–22 s matches gold (C).

L.2 Video-MME performance of our trained model

Figure 10 presents a qualitative example from the Video-MME short-video subset. The baseline

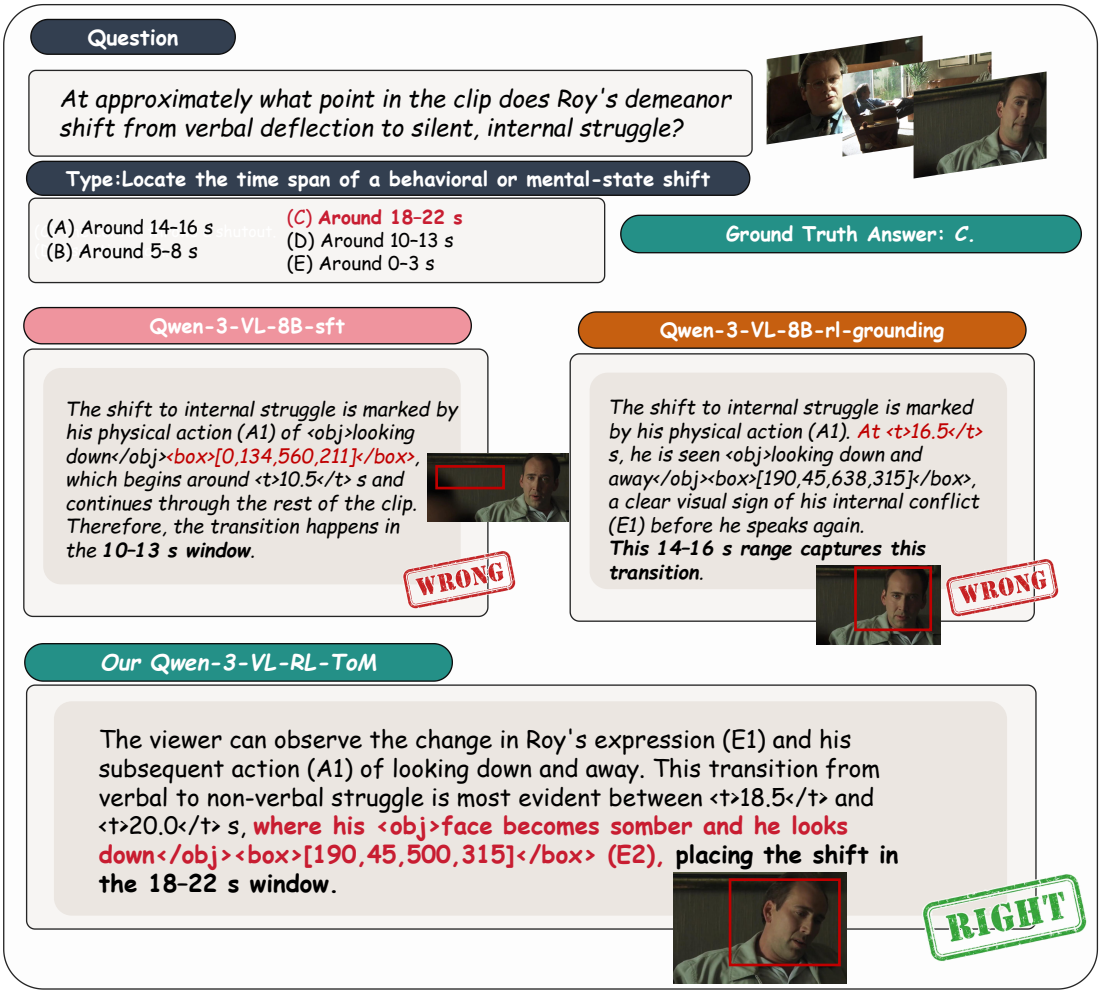

Figure 9: A hard G2 case from ToM-Video-val. SFT, RL-Grounding, and RL-ToM are asked to localize when Roy's demeanor shifts from *verbal deflection* to *silent, internal struggle* (gold answer C, 18–22 s clip-relative). Each panel shows the predicted frame, bounding box, and reasoning trace.

Qwen3-VL-8B-Instruct model incorrectly predicts the crowd is cheering for a layup-like event, relying primarily on a coarse understanding of the basketball scene. It does not explicitly connect the decisive shot event with the later crowd reaction. In contrast, our Qwen-VL-RL-ToM model predicts the correct answer, "three-point shoutout", and grounds its rationale in temporally separated evidence: the player shooting the ball at around 12.1 seconds, the crowd celebrating at around 22.0 seconds, and the announcer's cue. This example illustrates a concrete failure mode of the baseline and suggests that our trained model can better align its multiple-choice decision with localized visual-temporal evidence.

L.3 Social-IQ performance of our trained model (Detailed)

Analysis. Both models arrive at the correct answer (A), but through fundamentally different reasoning strategies. **Qwen3-Base** adopts an *elimination-by-exclusion* approach(Fig. 11): it catalogues the guest's visible laughter across the video, concludes that no option perfectly captures "amusement," and selects (A) as the least-wrong choice with an explicit caveat (*"despite its imperfect fit"*). The reasoning is logically sound but remains at the surface level of observable affect, without connecting the host's specific utterance to an internal mental state. **Qwen3-RL-ToM** instead reasons *causally* via grounded evidence(Fig.12): it identifies the mo-

Figure 10: Action reasoning example on Video-MME

ment the guest's expression *shifts* (E1→E2) in response to the host's utterance (L2), and infers the underlying belief state (B2: "he might be the target of a prank") that gives rise to the emotion of caution. This mirrors the theory-of-mind inference required by the question and produces a concise, evidence-anchored chain. Notably, the original evaluation mis-recorded the Base prediction as (D) due to generation truncation at 512 tokens (the response ended mid-sentence during the analysis of option (D)).

M Use of Large Language Models

We disclose the use of large language models (LLMs) in the four following aspects of this work.

(1) LLMs are integral to the automated annotation pipeline (§3): GPT-120b-OSS is used as the Film Selector. Gemini 2.5 Pro is used as the node-extractor and QA-generator, a local Qwen3-VL-8B is used for bounding-box annotation, and Qwen2.5-VL-72B is used for bbox quality inspection. The exact prompts of these stages are provided in Appendix C.

(2) Model training. The 8B vision–language model that we train (a Qwen3-VL backbone, depending on the experiment) is itself an LLM-based VLM.

(3) Baseline evaluation. We evaluate proprietary MLLMs through public APIs and open-source MLLMs locally, as listed in Section 5.

(4) Writing assistance. An LLM was used to polish English wording in selected paragraphs of this paper; all technical claims, equations, and ex-

Question

How does the person on the couch feel when the person with glasses says "you wouldn't know whether I was doing that to you"?

Type:Emotional Consequence

(A) **He feels cautious since it's possible he's being fooled.**
(B) The man feels proud of his collection of antique spoons.
……

Ground Truth Answer: A.

Social-IQ 2.0

Qwen-3-VL-8B-I

Based on the visual sequence provided, the person on the couch (Eddie Redmayne) reacts with increasing amusement and laughter to the host's (Stephen Colbert) comments.
In Frame 1, he is smiling.
In Frame 2, he is laughing while adjusting his tie.
In Frame 7, he is covering his mouth, likely laughing.
In Frame 8, he is laughing with his hand to his mouth.
In Frame 14, he is laughing and leaning forward.
In Frame 15, he is laughing while the host leans over his desk.
In Frame 16, he is still laughing.
The question asks how he feels when the host says, "you wouldn't know whether I was doing that to you?" This line is a classic comedic setup, implying the host is playing a trick or making a joke that the guest might not be able to discern. The guest's reaction is clearly one of amusement and laughter, indicating he finds the joke funny and is enjoying the interaction.
None of the options (A, B, C, D) directly describe the feeling of amusement or laughter. However, option (A) "He feels cautious since it's possible he's being fooled" is the closest to a plausible interpretation. The phrase "you wouldn't know whether I was doing that to you" implies a potential trick or deception. The guest's laughter suggests he is not actually being fooled or scared, but rather is enjoying the joke. However, the setup of the joke might make him momentarily consider the possibility of being tricked, which is a common reaction to such a line.
The other options (B, C, D) are completely unrelated to the context of the conversation and the guest's visible reactions.
Therefore, the best answer, despite its imperfect fit, is:
(A) He feels cautious since it's possible he's being fooled.
This is because the line "you wouldn't know whether I was doing that to you" inherently implies a potential trick or deception, and the guest's initial reaction might involve a moment of consideration or caution before he fully embraces the humor. His subsequent laughter indicates he finds the situation amusing, not frightening or confused.
<answer>A</answer>

Figure 11: Emotional Consequence On Social-IQ 2.0 (Base)

perimental numbers are produced and verified by the authors.

Question

How does the person on the couch feel when the person with glasses says "you wouldn't know whether I was doing that to you"?

Type:Emotional Consequence

(A) He feels cautious since it's possible he's being fooled.
(B) The man feels proud of his collection of antique spoons.
……

Ground Truth Answer: A.

Social-IQ 2.0

Our Qwen-VL-RL-ToM

<think>The viewer sees the man on the couch (Eddie) at <t>5.5</t>s <obj>smiling and looking down</obj><box>[0,22,208,358]</box> (E1). When the host (Stephen) says at <t>22.0</t>s, "you wouldn't know whether I was doing that to you" (L2), the man's expression shifts. He looks up with a <obj>slightly wary and thoughtful expression</obj><box>[131,20,464,358]</box> at <t>23.0</t>s (E2). This indicates he believes he might be the target of a prank (B2), causing a feeling of caution.</think>
<answer>A</answer>

RIGHT

EVIDENCE

EVIDENCE

Figure 12: Emotional Consequence On Social-IQ 2.0 (rl-tom)