StoryTR: Narrative-Centric Video Temporal Retrieval with Theory of Mind Reasoning

Anonymous ACL submission

Abstract

Current video moment retrieval excels at action-centric tasks but struggles with narrative content. Models can see *what is happening* but fail to reason *why it matters*. This semantic gap stems from the lack of **Theory of Mind (ToM)**: the cognitive ability to infer implicit intentions, mental states, and narrative causality from surface-level observations. We introduce **StoryTR**, the first video moment retrieval benchmark requiring ToM reasoning, comprising 8.1k samples from narrative short-form videos (shorts/reels). These videos present an ideal testbed. Their high information density encodes meaning through subtle multimodal cues. For instance, a glance paired with a sigh carries entirely different semantics than the glance alone. Yet multimodal perception alone is insufficient; ToM is required to decode that a character "smiling" may actually be "concealing hostility." To teach models this reasoning capability, we propose an **Agentic Data Pipeline** that generates training data with explicit three-tier ToM chains (intent decoding, narrative reasoning, boundary localization). Experiments reveal the severity of the reasoning gap: Gemini-3.0-Pro achieves only 0.53 Avg IoU on StoryTR. However, our 7B **Shorts-Moment** model, trained on ToM-guided data, improves +15.1% relative IoU over baselines, demonstrating that *narrative reasoning capability matters more than parameter scale*.

1 Introduction

Video moment retrieval (VMR) has achieved remarkable success on action-centric benchmarks (Lei et al., 2021; Gao et al., 2017; Hendricks et al., 2017), where queries like "find the person running" can be resolved through direct visual pattern matching. However, when confronted with *narrative-centric* content, current multimodal large language models (MLLMs) exhibit a critical **semantic gap**: they excel at recognizing *explicit* visual patterns but struggle to infer *implicit* narrative intents. In essence, *current models are skilled at seeing "what is happening" explicitly, but fail to reason "why it matters" implicitly*. This gap becomes pronounced when visual surfaces contradict internal mental states. A character may be smiling while plotting revenge, or appear calm while harboring deep grief.

Narrative short-form videos (shorts/reels) serve as the ideal testbed for this challenge. Unlike simple video clips, shorts are *high-density streams of social signals*. They compress complete story arcs into 1-3 minutes through dialogue reversals, sudden BGM shifts, and close-up micro-expressions. This extreme information density demands **native multimodal perception**: a protagonist's glance (visual) paired with a sigh (audio) carries entirely different semantics than the glance alone; a visually calm scene overlaid with dissonant music signals hidden tension. Yet multimodal perception alone is insufficient. Consider the query "Find the moment the protagonist realizes the betrayal." The visual surface may show only silence and a lingering gaze, while the audio reveals a subtle tremor in the voice. Multimodal models can *perceive* these cues but cannot *interpret* them: they see "a person pausing" but fail to infer "realization of betrayal." Bridging this gap requires **Theory of Mind (ToM)** (Premack and Woodruff, 1978). ToM is the cognitive ability to attribute mental states, beliefs, and intentions to others. It transforms scattered multimodal cues into coherent narrative understanding.

We hypothesize that ToM reasoning, while absent in current models, is *learnable through data*. Traditional VMR annotations provide mere timestamps without reasoning traces. They cannot teach models *why* a moment matters narratively. To address this, we propose an **Agentic Data Pipeline** that generates training data with explicit reasoning chains. The pipeline employs a *Clipper Agent* for fine-grained multimodal perception (capturing subtle visual and audio cues), and a *Self-QA Agent*

that produces answers guided by **three-tier ToM reasoning**. The three tiers are: (1) *Intent Decoding*, which infers character goals and mental states; (2) *Narrative Reasoning*, which traces causal chains and plot significance; and (3) *Boundary Localization*, which grounds abstract understanding to precise temporal segments. This approach distills ToM capabilities from advanced models (Gemini-3.0-Pro) into structured training data, enabling smaller models to acquire narrative reasoning through explicit chain-of-thought supervision.

Our experiments quantify the severity of this cognitive gap: even Gemini-3.0-Pro achieves only 0.53 Avg IoU on narrative retrieval, while action-optimized models like Qwen3-Omni collapse to 0.07. Crucially, our 7B Shorts-Moment model, trained on ToM-guided data, achieves +15.1% relative IoU improvement, outperforming larger models and demonstrating that *reasoning capability matters more than parameter scale*.

We make three contributions:

- **StoryTR Benchmark**: The first VMR benchmark requiring Theory of Mind reasoning, with 8.1k samples designed to test intent decoding, causal reasoning, and narrative understanding.
- **ToM-Guided Data Paradigm**: A principled approach demonstrating that explicit reasoning chains can transfer cognitive capabilities from large to small models, addressing the "invisible intent" problem in video understanding.
- **Empirical Validation**: Evidence that narrative reasoning is learnable. Our 7B model surpasses 30B+ baselines, proving that cognitive depth outweighs computational scale.

2 Related Work

2.1 Video Moment Retrieval: Success and Limitations

Video Moment Retrieval (VMR) aims to localize temporal segments corresponding to natural language queries. Early approaches framed this as cross-modal alignment (Gao et al., 2017; Hendricks et al., 2017), with benchmarks like QVHighlights (Lei et al., 2021) and TVR (Lei et al., 2020) driving progress on action-centric content.

The advent of multimodal large language models (MLLMs) has significantly advanced VMR capabilities. Models such as InternVideo2 (Wang et al., 2025), Qwen-VL (Bai et al., 2025; Team, 2025), and InternVL3 (Zhu et al., 2025) demonstrate strong cross-modal alignment, while native multimodal architectures like Gemini (Google DeepMind, 2025) and ARC-Hunyuan (Ge et al., 2025) process video, audio, and text as unified tokens.

The Perception-Cognition Gap. Despite these advances, current VMR systems share a fundamental limitation: they excel at *perceptual* tasks (detecting "when someone jumps") but struggle with *cognitive* tasks (understanding "why the jump matters narratively"). This gap manifests in two dimensions. First, *temporal duration*: benchmarks like Video-MME (Fu et al., 2024) and LongVideoBench (Wu et al., 2024) reveal that reasoning degrades with video length. Second, and more fundamentally, *semantic depth*: existing datasets lack queries requiring inference about character intent, emotional dynamics, or narrative causality. Our work addresses this second dimension by moving VMR from pattern matching to narrative reasoning.

2.2 Narrative Understanding and Theory of Mind

From Perception to Cognition. Existing narrative benchmarks like MovieNet (Huang et al., 2020) and ShotBench (Liu et al., 2025) evaluate *what* happens (shot boundaries, scene transitions) rather than *why* moments matter (character mental states, plot significance). Models can segment videos but cannot explain narrative function. Narrative short-form videos amplify this challenge by compressing story arcs into 1-3 minutes as *high-density signal streams*. Unlike long-form content where narrative cues are distributed across hours, every second in shorts carries critical story weight. A model may detect "a woman smiling" but fail to infer "she is concealing hostility," which requires reasoning about mental states invisible to pure perception.

ToM as Solution. Theory of Mind (ToM) (Premack and Woodruff, 1978; Wellman et al., 2001), the capacity to attribute mental states and intentions to others, has proven effective for dialogue understanding (Sap et al., 2022) and text-based story comprehension (Rashkin et al., 2018; Kosinski, 2023; Xie et al., 2022). ToM enables inferring what characters believe, what they intend, and why their actions matter within the narrative structure. Our work bridges this gap by introducing ToM reasoning to VMR, demonstrating that the "invisible intent" problem can be addressed through explicit cognitive reasoning rather than improved perception alone.

3 The StoryTR Benchmark

We construct StoryTR (Story-centric Temporal Retrieval), a benchmark explicitly designed to test **Theory of Mind reasoning** in video moment retrieval. Unlike action-centric benchmarks that evaluate perceptual capabilities ("when does someone run?"), StoryTR evaluates cognitive capabilities ("when does the protagonist realize the betrayal?"). This distinction is fundamental: the former requires pattern matching, the latter requires reasoning about mental states.

3.1 Why Short Dramas?

Design Rationale. We select narrative short-form videos (shorts/reels) not for novelty, but because they constitute an ideal testbed for the perception-cognition gap. Short dramas are *high-density streams of social signals*. They compress complete story arcs into 1-3 minutes through dialogue reversals, BGM shifts, and micro-expressions. This density creates a rigorous evaluation environment: every second carries narrative weight, and success requires understanding *why* moments matter, not just *what* happens.

Data Source. We collect videos featuring character development, emotional climaxes, and plot twists comparable to feature-length content. This compressed storytelling amplifies the ToM challenge. A protagonist's subtle glance may signal betrayal, jealousy, or reconciliation depending on narrative context. This information is invisible to perception-only models.

Scale and Quality. StoryTR comprises 8,141 samples (7,330 training, 811 testing). We prioritize *annotation depth* over scale: each query-answer pair is designed to require genuine narrative reasoning, with 811 test samples human-calibrated for rigorous evaluation (see Appendix A).

Data Consent and Usage Rights. All videos in StoryTR are sourced from publicly available short-form video platforms where content is released under commercial licenses permitting research use. We obtained explicit permission from content distributors for academic research purposes. Our dataset contains only metadata annotations (queries, timestamps, reasoning chains) rather than redistributing original video content, ensuring compliance with copyright and licensing requirements. Annotators were informed that their labels would be used for academic research and model training, with no personally identifiable information collected or retained.

Dataset	Domain	Video	Audio	Narrative
QVHighlights	Vlog/News	✓	✗	✗
Charades-STA	Activity	✓	✗	✗
TVR	TV Show	✓	✗	✗
ShortVID	Short Video	✓	✓	✗
MomentSeeker	Multi-domain	✓	✗	✓
StoryTR (Ours)	**Short Drama**	✓	✓	✓

Table 1: Comparison with existing VMR benchmarks. StoryTR is the only dataset combining full modality (video + audio) with narrative-focused queries requiring character intent and plot reasoning.

3.2 Query Types: A Cognitive Hierarchy

StoryTR queries are organized into three tiers of increasing cognitive complexity, designed to systematically probe the perception-cognition gap:

- **Tier 1: Intent Decoding** (First-order ToM). These queries require inferring the underlying purpose or motivation behind character actions beyond surface-level observations. Models must reason about *why* characters behave as they do, not just *what* they do. Example: "Find the explanation for why Charles is not present." Success requires understanding that absence implies a causal reason that must be explicitly stated or implied in dialogue or narration.
- **Tier 2: Narrative Reasoning** (Second-order ToM). These queries demand understanding story structure, plot progression, and causal logic across temporal sequences. Models must trace how narrative states evolve and identify pivotal moments where relationships, knowledge states, or power dynamics shift. Example: "Locate the moment when their relationship shifts from conflict to reconciliation." This requires recognizing abstract narrative transitions rather than concrete visual events.
- **Tier 3: Boundary Localization** (Evidence Grounding). These queries require precise temporal grounding of abstract narrative concepts to specific multimodal evidence. Models must identify the exact boundaries where implicit mental states become manifest through observable cues (dialogue, expressions, actions, music). Example: "Find evidence demonstrating the character's growing suspicion." Success requires aggregating subtle cues across modalities to pinpoint when abstract emotions crystallize into detectable signals.

The detailed reasoning process for generating these queries is described in Section 4.2.

3.3 Comparison with Existing Benchmarks

Table 1 positions StoryTR within the VMR landscape. The key distinction is not domain (short drama vs. vlog) but *cognitive requirement*. StoryTR is the only benchmark requiring Theory of Mind reasoning. Its queries probe character intent, emotional dynamics, and narrative causality rather than observable actions. Combined with native multimodal input (video + audio), StoryTR provides the first rigorous testbed for evaluating whether models can bridge the perception-cognition gap.

4 Methodology

Our methodology is grounded in a key insight: **ToM reasoning is learnable through data, not architecture**. Traditional VMR annotations provide only timestamps. They tell models *where* to look but not *why* a moment matters. This supervision gap explains why scaling model parameters alone fails to bridge the perception-cognition divide.

We address this through a **ToM-Guided Data Paradigm**: generating training data that contains explicit reasoning chains, enabling smaller models to acquire cognitive capabilities through chain-of-thought supervision. As illustrated in Figure 1, we implement this paradigm via an Agentic Data Pipeline with two components: a *Clipper Agent* that transforms raw video into structured perceptual logs, and a *Self-QA Agent* that synthesizes query-answer pairs with explicit three-tier ToM reasoning. The pipeline distills Gemini-3.0-Pro's narrative understanding into the open-source ARC-Hunyuan model.

4.1 Clipper Agent: Perception Foundation

The Clipper Agent transforms raw video into structured perceptual logs. These logs serve as the *sensory evidence* upon which ToM reasoning operates. This addresses a key challenge: ToM requires reasoning about cues that may exist only in non-visual modalities (e.g., a dissonant chord signaling hidden tension in a visually calm scene).

Multimodal Log Generation. Driven by Gemini-3.0-Pro's native multimodal capabilities, the agent produces three temporally-aligned streams. First, *Actions* capture fine-grained physical behaviors ("woman raises hand to face and wipes tears") with character identification via physical descriptors to avoid hallucination. Second, *Dialogue* records transcribed speech with speaker attribution. Third, *Sounds* include both diegetic audio (footsteps, door slams) and non-diegetic elements (BGM shifts). The third stream is critical for short dramas, which heavily rely on music to signal emotional tone invisible to visual analysis.

Shot-Aware Tracking. A key design principle is explicit re-identification across camera cuts. This prevents the common error of merging actions from different characters across shot boundaries. It ensures the logs provide accurate evidence for downstream reasoning.

The output is a dense, timestamped "screenplay" (Figure 1) that bridges raw multimodal signals to symbolic representations suitable for cognitive reasoning. Importantly, this perception layer is an *enabler*, not the contribution itself. It provides the evidence base upon which ToM reasoning operates.

4.2 Self-QA Agent: ToM Reasoning Engine

The Self-QA Agent is the core of our data paradigm. While the Clipper Agent captures *what* happens, the Self-QA Agent generates training data that teaches models *why* moments matter. This is the cognitive reasoning missing from traditional VMR annotations.

Three-Tier ToM Reasoning. The agent produces query-answer pairs with explicit reasoning chains that mirror human narrative understanding:

Tier 1: Intent Decoding (First-order ToM). Given a query like "Find the explanation for why Charles is not present," the agent identifies that this seeks narrative reasoning, not visual detection. The underlying intent is to justify a character's absence; key evidence markers include dialogue containing explanatory phrases. This tier teaches models to distinguish queries requiring cognitive reasoning from those requiring perception.

Tier 2: Narrative Reasoning (Second-order ToM). The agent anchors the query within story structure, reasoning that absence explanations typically appear early in a scene to manage audience expectations. This tier generates explicit chains explaining *why* a specific segment constitutes the correct match. It teaches models the causal logic of narrative construction.

Tier 3: Boundary Localization (Evidence Grounding). The agent grounds abstract concepts in concrete evidence: "At 00:17, the older woman states 'Charles had to go abroad for work,' provid-

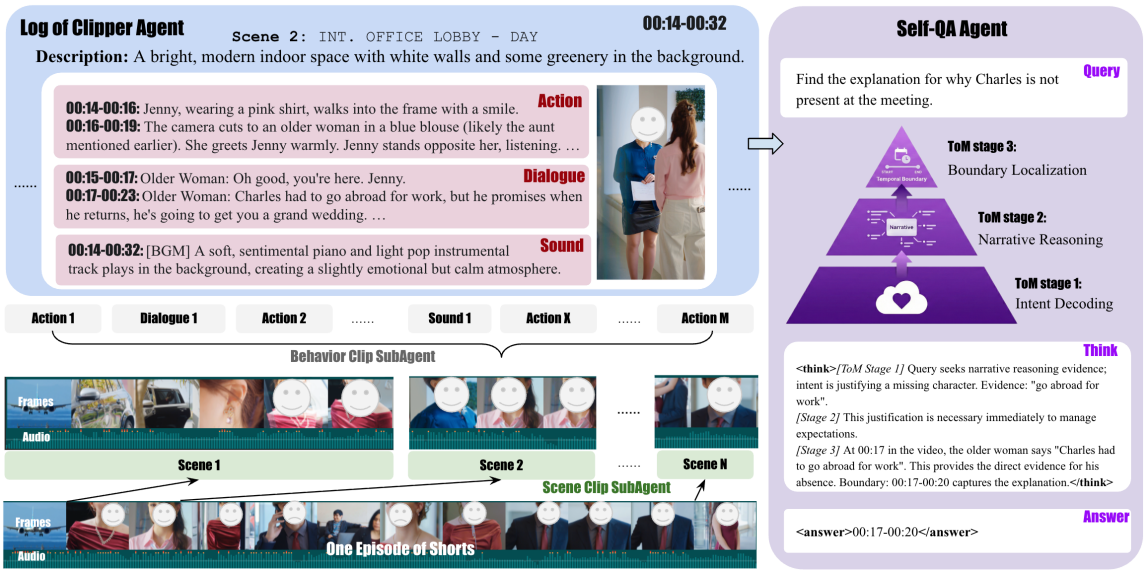


Figure 1: Overview of our native multimodal perception pipeline for narrative shorts (short dramas/reels). We leverage complementary foundation models: Gemini-3.0-Pro for high-quality annotation and ARC-Hunyuan for efficient feature extraction and training. The pipeline processes short drama videos through multimodal encoding to generate temporal localization results with detailed reasoning logs.

ing direct evidence." The boundary 00:17–00:20 captures the complete dialogue turn. This tier ensures models learn to retrieve segments grounded in explicit multimodal evidence while respecting narrative arc completeness.

Data as Curriculum. The resulting training data functions as a cognitive curriculum. Rather than learning implicit patterns from timestamps alone, models acquire explicit reasoning strategies through chain-of-thought supervision. This paradigm enables knowledge transfer by distilling Gemini-3.0-Pro's ToM capabilities into smaller, deployable models.

4.3 Knowledge Distillation via Supervised Fine-Tuning

Backbone Selection. We select ARC-Hunyuan (Ge et al., 2025) (7B parameters) as the student model. It is a unified video-language model with native audio processing via Whisper integration. The choice is deliberate: we aim to demonstrate that *reasoning capability can be transferred through data*, independent of model scale.

ToM Distillation. We perform supervised fine-tuning using the Self-QA Agent's outputs: query-answer pairs enriched with three-tier reasoning chains. This process distills Gemini-3.0-Pro's ToM capabilities into the 7B model, teaching it not just *where* to localize but *how* to reason about narrative structure. The resulting model, **Shorts-Moment**, acquires cognitive capabilities that perception-only training cannot provide.

5 Experiments

Our experiments test three hypotheses derived from our theoretical framework:

- **H1 (Perception-Cognition Gap)**: Current models, despite strong perceptual capabilities, fail on narrative retrieval. This validates that the gap exists.
- **H2 (ToM Learnability)**: Explicit ToM reasoning chains improve temporal precision. This validates that cognitive capabilities can be taught through data.
- **H3 (Reasoning $>$ Scale)**: A 7B model with ToM training outperforms larger models without it. This validates that cognitive depth matters more than parameter count.

5.1 Experimental Setup

We evaluate on the StoryTR test set (811 samples). Performance is measured using Intersection over Union (IoU) at thresholds $\theta \in \{0.3, 0.5, 0.8\}$, along with Precision and Recall.

Video and Audio Sampling. Following ARC-Hunyuan's protocol, we apply adaptive sampling based on video duration. For videos $\leq$150s, we sample at 1 FPS (one frame per second); for longer

Models	Precision				Recall				IoU			
	@0.3	@0.5	@0.8	AVG	@0.3	@0.5	@0.8	AVG	@0.3	@0.5	@0.8	AVG
Closed-source Models												
Gemini-3.0-Pro	0.804	0.698	0.474	0.659	0.866	0.795	0.551	0.718	0.757	0.580	0.247	0.532
Gemini-2.5-Pro	0.766	0.665	0.462	0.630	0.840	0.789	0.541	0.706	0.719	0.562	0.233	0.505
Gemini-2.5-Flash	0.394	0.365	0.276	0.344	0.369	0.310	0.181	0.293	0.331	0.245	0.085	0.231
Open-source Baselines												
Qwen3-Omni	0.152	0.127	0.089	0.130	0.127	0.098	0.047	0.099	0.110	0.065	0.013	0.074
ARC-Hunyuan(Ge et al., 2025)	0.551	0.457	0.281	0.447	0.647	0.599	0.406	0.542	0.487	0.353	0.127	0.344
Ours												
Shorts-Moment (Ours)	**0.631** (+14.5%)	**0.521** (+14.0%)	**0.309** (+10.0%)	**0.493** (+10.3%)	**0.698** (+7.9%)	**0.622** (+3.8%)	**0.420** (+3.4%)	**0.570** (+5.2%)	**0.572** (+17.5%)	**0.420** (+19.0%)	**0.158** (+24.4%)	**0.396** (+15.1%)

Table 2: Video Moment Retrieval Evaluation Results. We report accuracy at different thresholds (@0.3, @0.5, @0.8) and the mean (AVG). For our SFT model (Shorts-Moment), we additionally report the relative improvement over the baseline (ARC-Hunyuan) in parentheses. Best results in each category are marked in **bold**. We underline Gemini-3.0-Pro to denote its role as the teacher model for data generation, which also achieves the best performance among closed-source models.

videos, we uniformly sample 150 frames. Each frame is selected from the midpoint of its corresponding time segment. For audio, we use 16kHz sampling rate with a maximum of 150 segments. Each segment captures 2 seconds of audio; for videos exceeding 300s, we uniformly sample 150 segments across the timeline. Audio shorter than 1 second is zero-padded.

Input Format. Following ARC-Hunyuan(Ge et al., 2025)'s protocol, we construct inputs by concatenating N visual frame tokens with the text query, where each frame token encodes one sampled image. Audio features are extracted via the Whisper encoder and fused with visual tokens before being fed to the model.

Baselines. We compare against both closed-source and open-source models (see Appendix B for detailed specifications). For closed-source baselines, we evaluate the Gemini model family: Gemini-3.0-Pro serves as the teacher model for our data generation pipeline, representing Google's most advanced native multimodal architecture with 1M context length; Gemini-2.5-Pro provides a high-performance comparison point with advanced post-training via reinforcement learning; Gemini-2.5-Flash serves as an efficiency baseline as a lightweight distilled variant. For open-source baselines, we evaluate Qwen3-Omni, a 30B Mixture-of-Experts model capable of end-to-end omni-modal processing, and ARC-Hunyuan, a dense 7B model with timestamp overlay mechanism designed for temporal localization, which serves as the backbone for our Shorts-Moment model.

5.2 H1: The Perception-Cognition Gap Exists

Table 2 quantifies the severity of the gap. Despite strong performance on action-centric benchmarks, models collapse on narrative retrieval:

Perception Alone Fails. Qwen3-Omni (30B MoE) achieves only 0.074 Avg IoU. Despite state-of-the-art perceptual capabilities, this model fails catastrophically when queries require reasoning about intent. Gemini-2.5-Flash, optimized for efficiency, drops to 0.231. Even Gemini-3.0-Pro, the strongest baseline, achieves only 0.532 Avg IoU.

The Gap Is Cognitive, Not Perceptual. These models can detect "a woman smiling" but cannot infer "she is concealing hostility." StoryTR's Tier 2-3 queries (60% of test set) require exactly this cognitive leap. This explains why perception-only models fail despite adequate sensory processing.

5.3 H2: ToM Reasoning Is Learnable

Precision at Strict Thresholds. The most striking result appears at IoU@0.8: our model improves +24.4% over the baseline. This validates that ToM reasoning enables *precise* boundary localization. The three-tier process (intent $\rightarrow$ narrative $\rightarrow$ evidence) helps models pinpoint exact narrative boundaries rather than approximate vicinities.

Explicit Chains Outperform Implicit Learning. Base ARC-Hunyuan (trained on timestamps only) achieves 0.344 Avg IoU. It identifies general locations but lacks precision. Our ToM-enhanced model achieves 0.396 (+15.1%), demonstrating that explicit reasoning chains teach models *why* boundaries matter, not just *where* they are.

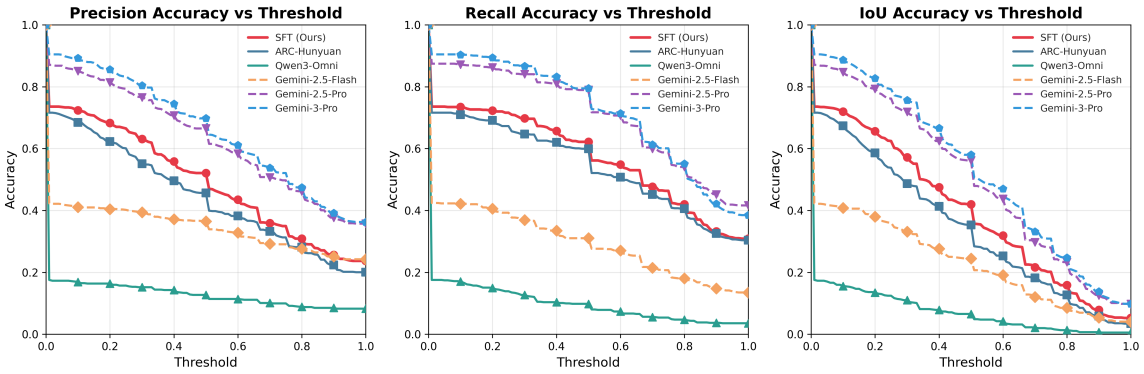


Figure 2: Accuracy results for curves

5.4 H3: Reasoning Capability > Parameter Scale

7B Beats 30B. Our 7B Shorts-Moment outperforms the 30B Qwen3-Omni by 5× in Avg IoU (0.396 vs 0.074). This validates our core thesis: cognitive depth matters more than computational scale. A model *taught to reason* outperforms a larger model that merely perceives.

Approaching Closed-Source SOTA. Our model surpasses Gemini-2.5-Flash (0.396 vs 0.231 Avg IoU) and narrows the gap to Gemini-3.0-Pro (0.396 vs 0.532). It achieves 74% of the teacher's performance with a fraction of parameters. This validates that ToM capabilities can be effectively distilled through our data paradigm.

6 Discussion and Analysis

6.1 Threshold Sensitivity Analysis

Figure 2 presents accuracy curves across varying IoU thresholds, revealing model robustness for temporal localization. As thresholds increase from 0.0 to 1.0, all models degrade, but at different rates.

Performance Stability. Our model demonstrates superior stability in the mid-range region (0.3-0.6). In Precision, we maintain >50% accuracy until threshold 0.5, while ARC-Hunyuan drops faster. Similarly, Recall shows sustained higher rates across all thresholds, indicating improved coverage of narrative moments. It suggests that ToM-guided training teaches models to identify *semantically coherent* boundaries rather than arbitrary cut points, since narrative beats naturally align with specific temporal spans.

Temporal Localization Robustness. Our model exhibits graceful degradation compared to open-source baselines. While Qwen3-Omni collapses at moderate thresholds, we maintain meaningful IoU scores up to 0.7 through narrative-aware fine-tuning that respects story boundaries. The divergence is especially pronounced in the IoU curve. The gap widens as thresholds increase, indicating that ToM reasoning provides the most value when *precise* temporal grounding is required. This is exactly the regime where understanding "why" a moment matters becomes critical for determining "where" it begins and ends.

Comparison with Closed-Source Models. Gemini-2.5-Flash declines steeply after threshold 0.3, while Gemini-2.5-Pro and 3.0-Pro maintain stability at the cost of larger sizes. Our 7B model closely tracks these larger models in Recall while offering deployment advantages. Notably, we intersect with Gemini-2.5-Flash at threshold 0.4 in Precision and maintain superiority thereafter. This demonstrates that specialized training enables better discrimination at stricter requirements, where narrative understanding matters most.

6.2 Case Study

Figure 3 crystallizes the perception-cognition gap. The query "*When is the DNA test result visually revealed to the group?*" appears simple but encodes a critical semantic distinction. *Revealed* implies **intentional disclosure**, a social act, not mere visual presence. The ground truth spans 00:24–00:28: the woman's deliberate flip (0:24), the document close-up (0:25), and the man's shocked reaction (0:28). These form a complete *communicative arc*.

Perception-Only Models Fail. ARC-Hunyuan predicts 00:25–00:27 (IoU=0.5), starting one second late because it anchors on the camera zoom-in rather than the character's initiating action. Its reasoning, *"camera zooms in showing the probability,"* reveals pure visual detection without intent modeling. It sees the document but misses the *act*

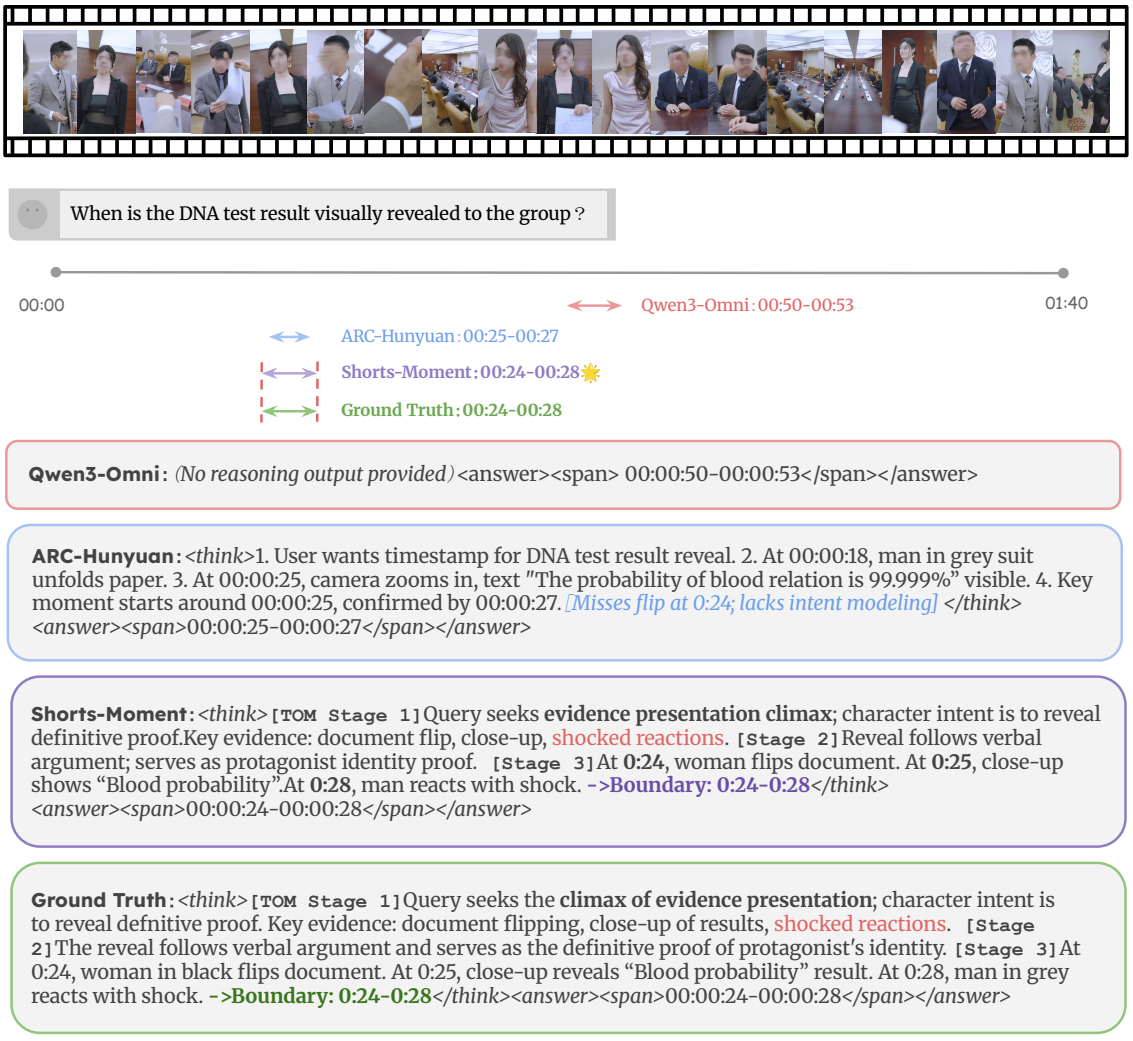


Figure 3: Case study comparing reasoning outputs for query "*When is the DNA test result visually revealed to the group?*". Shorts-Moment achieves exact match (IoU=1.0) via three-stage ToM reasoning, while ARC-Hunyuan misses the boundary (IoU=0.5) and Qwen3-Omni fails completely (IoU=0.0).

of revealing. Qwen3-Omni fails catastrophically (IoU=0.0), predicting 00:50–00:53. This 26-second offset demonstrates that parameter scale cannot substitute for cognitive capability.

ToM Reasoning Succeeds. Shorts-Moment achieves exact match (IoU=1.0) by reasoning through all three tiers. It identifies that the reveal *begins* at 0:24 when the woman flips the document. It understands this initiating action is semantically part of "revealing," not preparation. This requires *intent modeling*: the flip is deliberate disclosure, not incidental handling. The model extends to 0:28 to capture the shocked reaction, completing the narrative beat.

Generalization. This pattern generalizes: queries involving social verbs (*reveal, confront, comfort, deceive*) encode ToM concepts invisible to perception. ToM reasoning transforms temporal localization from "when is X visible?" to "when does X achieve its *narrative purpose*?"

7 Conclusion

This work addresses a fundamental gap in video understanding: current models can perceive *what is happening* but fail to reason *why it matters*. **StoryTR** provides the first rigorous testbed for narrative video moment retrieval, with queries explicitly designed to probe ToM capabilities across three cognitive tiers. Further, our **ToM-Guided Data Paradigm** shows that explicit reasoning chains can transfer cognitive capabilities from large to small models. It addresses the "invisible intent" problem through chain-of-thought supervision rather than parameter scaling. Our experiments validate a key insight for the field: **reasoning capability matters more than scale**.

Limitations

ToM Complexity. Our three-tier ToM framework captures first- and second-order reasoning but may not fully address higher-order ToM (e.g., "A believes that B believes that C intends..."). Extending to deeper recursive reasoning remains future work.

Cultural Variation. ToM reasoning is culturally situated. Narrative conventions differ across traditions. Our dataset (55% Chinese, 45% English) may not generalize to all storytelling cultures without adaptation.

Teacher Dependency. The data paradigm relies on Gemini-3.0-Pro for reasoning chain generation. While we demonstrate successful distillation to 7B models, the initial data creation requires access to frontier models.

Benchmark Scale. StoryTR contains 8.1k samples. This is sufficient to validate our hypotheses but smaller than million-scale datasets. Scaling while maintaining annotation depth is an open challenge.

Ethics Statement

This research involves video content analysis and human annotation, raising several ethical considerations that we have carefully addressed:

Data Collection and Consent. All videos used in StoryTR are legally obtained from publicly available platforms with proper licensing for research use. We secured explicit permission from content distributors for academic research purposes. We do not distribute copyrighted video content—only metadata annotations (queries, timestamps, reasoning chains)—ensuring compliance with copyright and fair use principles.

Privacy. Our dataset contains only commercially released content with no private or sensitive personal information. All content creators are professional actors performing in scripted productions. No surveillance footage, private recordings, or nonconsensual content is included.

Annotator Welfare. Human annotation was conducted by trained researchers as part of their academic responsibilities. No crowdworkers were employed. Annotators were fully informed about the research purpose and provided consent for their annotations to be used in published datasets.

Potential Misuse. While our technology enables better video understanding for applications such as content retrieval and accessibility, it could potentially be misused for unauthorized content analysis or surveillance. We advocate for responsible use within legal and ethical boundaries, and recommend that future applications implement appropriate consent and transparency mechanisms.

Bias and Representation. Our dataset may reflect biases present in commercial short drama content, including potential under-representation of certain demographics, cultures, or narrative styles. Future work should address representation across diverse narratives, cultures, and production contexts. We encourage researchers using StoryTR to consider these limitations when deploying models in real-world applications.

Dual Use. We acknowledge that improved video understanding capabilities could have dual-use implications. We commit to responsible disclosure and encourage the community to develop appropriate safeguards for deployment.

References

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. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2407.12679*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. 2025. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*.

Google DeepMind. 2025. A new era of intelligence with gemini 3. Blog post. https://blog.google/products/gemini/gemini-3/#gemini-3.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. *arXiv preprint arXiv:1708.01641*.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021. Qvhighlights: detecting moments and highlights in videos via natural language queries. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, page NIPS '21. Curran Associates Inc.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, Yu Qiao, Wanli Ouyang, Shengjie Zhao, and Ziwei Liu. 2025. Shotbench: Expert-level cinematic understanding in vision-language models. *arXiv preprint arXiv:2506.21356*.

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

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *EMNLP*.

Qwen Team. 2025. Qwen3-omni: A unified multimodal large language model. *arXiv preprint arXiv:2509.17765*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and et al. 2025. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, 72(3):655–684.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Yuqiang Xie, Yue Hu, Wei Peng, Guanqun Bi, and Luxi Xing. 2022. Comma: Modeling relationship among motivations, emotions and actions in language-based human activities. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3632–3644.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2507.12679*.

A Manual Annotation for Ground Truth Verification

A.1 Annotation Setup

To establish ground truth labels for evaluating our framework, we randomly selected 811 query-answer pairs (approximately 10% of StoryTR) stratified across Chinese and English short dramas. The annotation was conducted by the authors using Label Studio [1], an open-source platform

As shown in Figure 5, the interface displays the video on the left and StoryTR entry data on the right, including: `video_id` (drama + episode identifier), `query` (question), `timestamp` (model prediction), `usage` (query type), and `think` (model reasoning). Annotators completed two fields: (1) **Timestamp Accuracy** (binary: 1 if prediction is within 2 seconds of ideal boundaries; 0 otherwise), and (2) **Ground Truth** (precise temporal boundaries in MM:SS-MM:SS format, or `bad` for invalid queries).

A.2 Task Description and Annotation Protocol

Annotators followed a standardized protocol: (1) read query and understand target moment, (2) watch video and navigate to predicted timestamp, (3) label accuracy (1 if within $\leq$ 2s tolerance; 0 otherwise), and (4) provide ground truth (precise MM:SS-MM:SS boundaries for valid queries; `bad` for invalid queries). Figure 5 presents detailed task instructions and a complete annotation example demonstrating the process.

Category	Samples	Accuracy (%)
Overall	811	92.11
By Language		
Chinese	480	90.62
English	311	94.26
By Query Type		
Direct Localization	276	90.94
Event Identification	273	91.21
Evidence Localization	265	94.27

Table 3: Timestamp accuracy across categories. Accuracy is defined as predictions within 2 seconds of ground truth boundaries.

[1] https://github.com/HumanSignal/label-studio

A.3 Annotator Recruitment and Compensation

Recruitment. Manual annotation was conducted by the authors, who are graduate students and researchers with expertise in video understanding and natural language processing. All annotators are fluent in both English and Chinese, ensuring accurate evaluation of bilingual content in StoryTR.

Training and Qualification. Before formal annotation, annotators completed a training session using 50 practice samples to establish inter-annotator agreement on boundary definitions and quality criteria. Disagreements were resolved through discussion to ensure consistent annotation standards.

Compensation. As the annotation was performed by the research team as part of their academic responsibilities, no additional monetary compensation was provided beyond standard research assistantship stipends. The annotation workload (811 samples) was distributed across the team and completed over approximately 40 person-hours. This approach ensured high-quality annotations from domain experts while avoiding potential quality issues associated with crowdsourced annotation for tasks requiring deep narrative understanding.

Demographic Context. All annotators are based in research institutions in China, where the standard compensation for research assistants is commensurate with local cost of living and academic norms.

A.4 Results and Analysis

Table 3 presents accuracy results. Overall accuracy is **92.11%**, with Chinese short dramas achieving **90.62%** and English achieving **94.26%**. By query type, Direct Localization achieves **90.94%**, Event Identification **91.21%**, and Evidence Localization **94.27%**.

B Baseline Implementation Details

B.1 Experimental Setup

To ensure fair comparison across diverse architectures, all baseline experiments were conducted under identical conditions. All models were evaluated in a zero-shot setting with temperature $\tau = 1.0$ to standardize generation randomness. Video inputs were standardized at 2 FPS for all multimodal models. Identical prompt templates were used across all models, with minor adaptations only for specific chat formats (e.g., ChatML for Qwen).

```
You are a video time positioning assistant.  You need to locate the
corresponding time segment in the video based on the given query.

Video total duration:  {self.duration}

Query:  {self.query}

Please follow these rules when responding:
1.  Analyze the query within the <think> tag to understand the specific
information to locate.
2.  Provide the precise time segment within the <answer> tag, marking
timestamps with <span> tags.
3.  Format the output inside <answer> tags, converting all time ranges to the
pattern <span>HH:MM:SS - HH:MM:SS</span>.
4.  If the query content does not exist in the video, clearly state this.

Generate the response:
```

Figure 4: **System Prompt for Video Moment Retrieval.** The figure illustrates the structured instructions provided to the model for temporal grounding tasks. The prompt enforces a specific output format using XML-style tags for easy parsing. Variables enclosed in curly braces (e.g., {self.query}) are dynamically replaced during inference.

Model	API/Version	Developer	Release	Context	Architecture
Gemini-3.0-Pro	gemini-3.0-pro-preview	Google	Nov 2025	1M	Native multimodal; SOTA
Gemini-2.5-Pro	gemini-2.5-pro	Google	Mar 2025	1M	Native multimodal
Gemini-2.5-Flash	gemini-2.5-flash	Google	Apr 2025	1M	Distilled model
Qwen3-Omni	Qwen3-Omni-30B-A3B-Instruct	Alibaba	Sep 2025	32K	MoE (30B total)
ARC-Hunyuan	ARC-Hunyuan-Video-7B	Tencent	Jul 2025	20K	Dense 7B
Shorts-Moment (Ours)	Fine-tuned from ARC-Hunyuan	-	-	20K	Dense 7B + ToM

Table 4: **Model Specifications.** API/Version column shows the exact model identifier used in experiments. Shorts-Moment is our model fine-tuned from ARC-Hunyuan on StoryTR training set.

Evaluations were performed on NVIDIA A800 (80GB) GPUs. Closed-source models were accessed via Google Vertex AI. Open-source models were deployed using vLLM (ARC-Hunyuan) and HuggingFace Transformers (Qwen3-Omni) with KV cache optimization.

B.2 Model Specifications

We compare our approach against five state-of-the-art multimodal baselines. Detailed specifications are summarized in Table 4.

Gemini Model Family (accessed via Google Vertex AI):

- **Gemini-3.0-Pro** (API: gemini-3.0-pro-preview): Google's most advanced multimodal model. It features native support for text, video, and audio. In our experiments, it serves as the *Teacher model* for data generation.
- **Gemini-2.5-Pro** (API: gemini-2.5-pro): A high-performance model utilizing advanced post-training techniques including reinforcement learning.
- **Gemini-2.5-Flash** (API: gemini-2.5-flash): A lightweight model designed for efficiency, serving as an efficiency baseline.

Open-Source Baselines:

- **Qwen3-Omni** (Version: Qwen3-Omni-30B-A3B-Instruct): An omni-modal MoE model capable of end-to-end processing of text, audio, and video.
- **ARC-Hunyuan** (Version: ARC-Hunyuan-Video-7B): A dense 7B model with timestamp overlay mechanism for temporal localization. Serves as the backbone for our Shorts-Moment model.

Our Model:

- **Shorts-Moment**: Fine-tuned from ARC-Hunyuan-Video-7B on StoryTR training set. Acquires ToM reasoning and narrative moment retrieval capabilities through our Agentic Data Pipeline.

B.3 Prompt Strategy

To mitigate the influence of prompt engineering on evaluation metrics and ensure a strictly fair comparison, we employed a unified prompt template across all models. As illustrated in Figure 4, the identical instruction was fed to both the closed-source Gemini family and the open-source baselines (Qwen and Hunyuan) without any model-specific optimization. The prompt explicitly constraints the output format, requiring models to predict precise start and end timestamps alongside a reasoning rationale. By freezing the input instruction, we ensure that the observed performance variances are solely attributable to the intrinsic video understanding and reasoning capabilities of the respective architectures.

Annotation Instruction

Task Description

OBJECTIVE:

For each StoryTR query-answer pair, verify the predicted timestamp accuracy and provide ground truth labels.

ANNOTATION STEPS:

1. CAREFULLY READ THE QUERY

   Understand what narrative moment is being requested.

2. WATCH THE WHOLE VIDEO

   - Play the entire video to understand full context
   - Navigate to the predicted timestamp to evaluate correctness

3. LABEL TIMESTAMP ACCURACY (Boolean: 1 or 0)

   - Label = 1 (CORRECT): Temporal boundaries differ by ≤2 seconds AND capture the core narrative moment
   - Label = 0 (INCORRECT): Off by >2 seconds, wrong moment, or misses key narrative beats

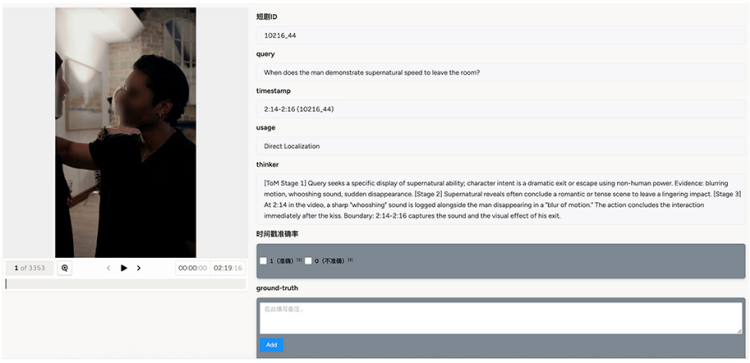
4. PROVIDE GROUND TRUTH

   - Valid query: Enter precise temporal boundaries in MM:SS-MM:SS format (e.g., 2:14-2:17) that capture the complete emotional arc
   - Invalid query: Enter "bad" if query references non-existent content is ambiguous, or violates zero-hallucination principle

KEY GUIDELINES:

- Focus on emotionally complete moments (include setup + climax + resolution)
- Prioritize emotional arc completeness over strict boundary precision
- For borderline cases (≤2s difference), consider if core moment is captured

Example



ANNOTATION PROCESS:

Step 1: Watch entire video (2:19 duration) for context

Step 2: Navigate to predicted timestamp 2:14-2:16

Step 3: Observe that 2:14-2:17 captures complete supernatural exit:

- 2:14: Whooshing sound begins
- 2:15: Blur motion effect visible
- 2:17: Man completely vanished, room empty (closure)

Step 4: Evaluate accuracy

- Prediction: 2:14-2:16 - Ground Truth: 2:14-2:17
- Difference: 1 second at end point (within 2s tolerance) - Core moment captured: ✅

ANNOTATION OUTPUT:

Timestamp Accuracy: 1 (CORRECT - within 2s tolerance)

Ground Truth: 2:14 - 2:17

RATIONALE: The predicted timestamp substantially captures the target moment. Though it ends 1 second early, the difference is within the 2s tolerance and the core supernatural exit is fully represented.

Figure 5: Annotation interface and task description. **Top**: Complete task description including annotation steps, guidelines.**Bottom**: Label Studio interface showing video player (left) and StoryTR entry fields (right) with annotation inputs, and a worked example demonstrating the labeling process for a supernatural speed query.