

Task-Oriented Moment Localization for Long Video Understanding with Near-Second Coverage

Xiangrui Liu^{1†} Minghao Qin^{2†} Yan Shu^{4†} Zhengyang Liang⁵ Yang Tian³

Chen Jason Zhang^{1✉} Bo Zhao^{3✉} Zheng Liu^{2✉}

¹Hong Kong Polytechnic University ²Beijing Academy of Artificial Intelligence

³School of AI, Shanghai Jiao Tong University ⁴University of Trento

⁵Singapore Management University

Abstract

Long-video understanding is increasingly framed as an agentic process that iteratively localizes task-critical moments before perception and reasoning. Moment localization is therefore a key bottleneck, and to address it we present **EviScan**, an evidence-seeking moment localization model built on an MLLM, designed for *task-oriented localization* and *near-second temporal scanning* on hour-long videos. To learn task-oriented localization, we construct **ToMo-Pile**, a training set that pairs temporal intervals with rationale traces explaining why the interval resolves the task. To make near-second scanning feasible on long videos, EviScan uses **DualKV Prefill** for memory-efficient streaming inference, and **Progressive Localization** to first narrow down candidate intervals and then refine boundaries using high-resolution evidence. We further introduce **ToMo-Bench**, a benchmark for task-oriented moment localization across diverse long-video scenarios. EviScan consistently outperforms specialized localizers and general-purpose MLLMs on both ToMo-Bench and standard temporal grounding benchmarks. When used as a plug-in localization module, it yields substantial gains on downstream long-video understanding benchmarks (MLVU and LongVideoBench), and these gains are larger than those brought by other moment-localization methods. Our code is open-sourced at: <https://github.com/lxrrrrrr/EviScan>

1 Introduction

As video understanding research moves from short clips to hour-long videos (Shu et al., 2024; Liu et al., 2025; Wang et al., 2025b; Shen et al., 2024; Zhang et al., 2024; Liang et al., 2025), an increasing number of studies cast video understanding

as an agentic, tool-augmented process (Yu et al., 2025; Tang et al., 2025; Yao et al., 2025; Wang et al., 2025a). Unlike traditional one-pass approaches (Liu et al., 2023; Li et al., 2023a,b; Yang et al., 2024), this paradigm iteratively *localizes* task-critical moments (i.e., temporal intervals) and *perceives* the corresponding evidence as required by the task until completion. In this setting, a precise task-oriented localization module is indispensable: without accurate localization, downstream perception and reasoning remain bottlenecked regardless of model capacity.

However, most existing studies typically adopt an off-the-shelf retriever or temporal grounding model for moment localization. These modules are not well suited to long-video understanding tasks. Retrievers rely primarily on surface-level semantics. Temporal grounding models, limited by context length, use sparse sampling that can easily skip decisive moments, causing the correct interval to be missed before localization even starts.

We argue that an effective localization module should satisfy two criteria: (1) **Task-oriented localization**: it should localize the moment that contains task-critical evidence, rather than merely matching surface semantics. For example, when the task asks "who broke the glass," the model should not merely retrieve frames showing "the glass is broken," but should localize the moment that causes the glass to break. (2) **Near-second temporal coverage**: it should scan hour-long videos at near-second temporal granularity to avoid missing decisive moments, while remaining computationally efficient.

Therefore, to meet the two desiderata above—task-oriented localization and near-second temporal coverage—we propose **EviScan (Evidence Scanning)**, a moment localization model built upon an MLLM, trained with our newly constructed dataset ToMo-Pile. Unlike existing temporal grounding datasets, **ToMo-Pile** is explicitly designed to foster task-oriented

† Equal contribution, ✉ Corresponding author.
xl77485@gmail.com, jason-c.zhang@polyu.edu.hk,
bo.zhao@sjtu.edu.cn, zhengliu1026@gmail.com.

moment localization. Beyond providing temporal interval annotations, it also includes reasoning traces that explain why a specific moment should be selected. ToMo-Pile encourages the model to align moment localization with task requirements rather than surface-level semantic similarity.

To achieve *near-second temporal coverage* over hour-long videos, we introduce an efficient streaming KV-prefill mechanism, **DualKV Prefill**. The video is processed segment by segment: for each segment, the model computes high-resolution *detailed KV caches* and then temporally downsamples them into compact *holistic KV caches*. This yields a two-level video representation, where detailed caches preserve local evidence and holistic caches provide lightweight long-range summaries. By accumulating holistic caches as historical memory and letting each new segment attend to them, DualKV Prefill injects long-range context without loading the full high-resolution sequence into a single context window, enabling near-second scanning with feasible compute and memory.

However, memory efficiency alone is not enough: near-second scanning still brings many irrelevant visual tokens, diluting attention and expanding the temporal search space. To address this, we propose **Progressive Localization**, a coarse-to-fine, multi-round strategy built on the two-level video representation. In the coarse stage, EviScan reasons over holistic KV caches to propose the task-aware candidate interval. In the fine stage, it selectively reloads the corresponding detailed KV caches and refines the boundaries to produce the final prediction. Combined with DualKV Prefill, this yields high-recall, near-second moment localization in long videos while focusing computation and attention on high-value evidence.

Moreover, existing temporal grounding benchmarks are primarily designed for conventional settings (e.g., phrase-to-moment matching) and therefore fail to reflect the localization demands of diverse long-video understanding (LVU) tasks. To bridge this gap, we introduce **ToMo-Bench**, a testbed for *task-oriented* moment localization in realistic LVU scenarios. It challenges models to identify key temporal intervals across 12 task categories and 32 video domains, spanning videos from a few seconds to over an hour. Each instance is built via a human-machine collaborative annotation pipeline to ensure high-quality and reliable temporal supervision.

In experiments, we benchmark EviScan against

specialized localizers and general-purpose MLLMs on our ToMo-Bench and three classic temporal grounding benchmarks: Charades-STA (Sigurdsson et al., 2016), ActivityNet (Caba Heilbron et al., 2015), and V-STaR (Cheng et al., 2025). EviScan consistently achieves stronger moment localization performance across all benchmarks. Moreover, EviScan also boosts downstream long-video understanding, outperforming tool-augmented video agents such as Video-Zoomer (Ding et al., 2025) and Conan (Ouyang et al., 2025).

Our contributions are three-fold: (1) We present EviScan, a task-oriented moment localization model for long videos that enables near-second temporal coverage with feasible computation. (2) We construct ToMo-Pile and introduce ToMo-Bench, providing training supervision and evaluation tailored to task-oriented localization across diverse LVU scenarios. (3) We demonstrate consistent improvements over strong specialized localizers and general MLLMs on both localization benchmarks and downstream LVU tasks.

2 Related work

2.1 Long Video Understanding

Recently, long video understanding has advanced rapidly, with powerful MLLMs extending context length and improving multi-step understanding (Li et al., 2025b; Qin et al., 2025; Yang et al., 2026). Despite this progress, existing systems still face two persistent challenges: high computational cost and the “lost-in-the-middle” effect, where critical evidence in long sequences is easily overlooked.

More recently, a complementary agentic paradigm has emerged, which combines LLM reasoning with active tool use for long-form video analysis. Representative methods such as VideoAgent (Wang et al., 2024) model the process as sequential information gathering and decision making over segmented clips, improving efficiency and interpretability for long-form VideoQA.

However, most prior work emphasizes agent planning while relying on fixed off-the-shelf retrieval backbones (e.g., CLIP (Radford et al., 2021) and LanguageBind (Zhu et al., 2024)). In long videos, retrieval quality is a first-order bottleneck, which motivates our EviScan that focuses on stronger task-relevant moment retrieval for agentic long-video understanding.

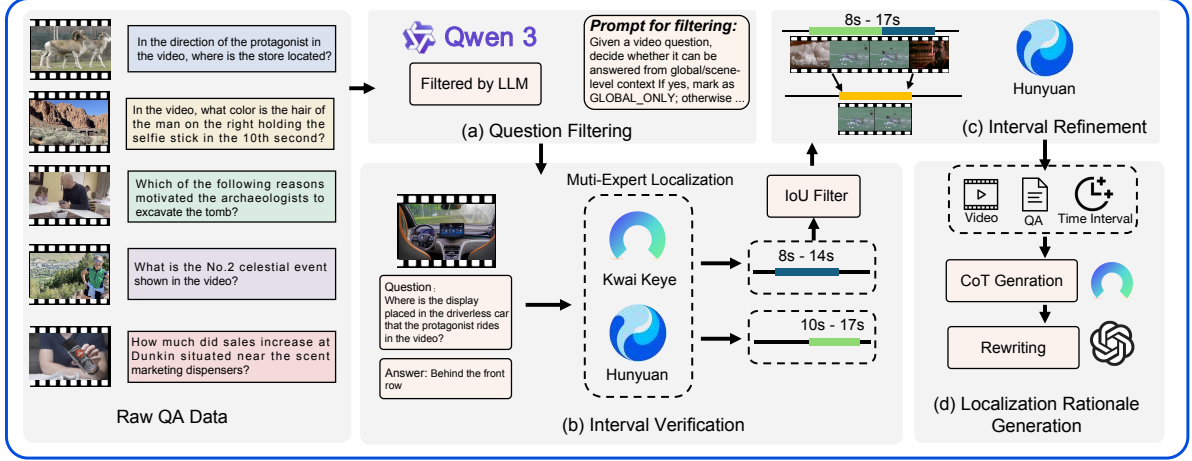


Figure 1: Overview of ToMo-Pile. We established multiple filtering mechanisms to obtain a high-quality Chain-of-Thought (CoT) dataset.

2.2 Moment Localization

Prior work on moment localization broadly falls into two lines: similarity-based retrieval and temporal grounding.

Retrieval-based methods select frames/clips via query–video similarity, often using pretrained vision–language models such as CLIP (Radford et al., 2021). In addition, there are also some methods explore specific similarity-based approaches. For instance, Flexible Frame Selector (Buch et al., 2025) introduces a [CTRL] register token to adaptively decide both how many frames to select and which frames to pick. However, these approaches struggle with deeper understanding for complex tasks.

Temporal grounding models aim to predict the start/end boundaries of the queried moment. Recent studies increasingly build on MLLMs and temporal-aware designs. For example, TimeChat (Ren et al., 2024) introduces a time-aware frame encoder that binds visual tokens with their corresponding timestamps at the frame level for temporal grounding. Time-R1 (Wang et al., 2025c) employs a reasoning-guided post-training framework with reinforcement learning and verifiable rewards to improve grounding accuracy. Recent generic MLLMs (e.g., Qwen2.5-VL (Bai et al., 2025), Key-VL-1.5 (Yang et al., 2025)) have also demonstrated certain capabilities for temporal grounding. However, localizing small intervals in long videos remains challenging due to their limited context window.

Motivated by these limitations, we propose EviScan for task-oriented moment localization in long videos with near-second temporal coverage.

3 Method

3.1 EviScan

EviScan is built on an MLLM backbone and trained on ToMo-Pile, enabling it to capture localization requirements across diverse LVU tasks. In addition, by integrating DualKV Prefill and Progressive Localization, EviScan achieves both high localization precision and efficient inference on long videos.

3.2 ToMo-Pile Construction

In many long-video understanding tasks, the target moment cannot be reliably identified by simple lexical matching between query words and salient visual entities. To this end, we construct **ToMo-Pile**, in which each sample contains both temporal intervals and localization rationales. We curate subsets from two video QA datasets, VideoR1 (Feng et al., 2025) and FineVideo (Farré et al., 2024). Together, they cover diverse video categories and question types.

Construction pipeline. As illustrated in Figure 1, we build our training data through a three-stage pipeline:

(1) Filtering. We first use an LLM to filter out questions that can be answered from global, scene-level cues alone. This step is motivated by the construction of VideoR1 and FineVideo, where a subset of questions follow obvious holistic patterns (e.g., asking about the overall setting of the video).

(2) Interval verification and refinement. For each remaining sample, we localize the evidence moment using a multi-expert procedure. Specifically, given the video and the ground-truth answer, two

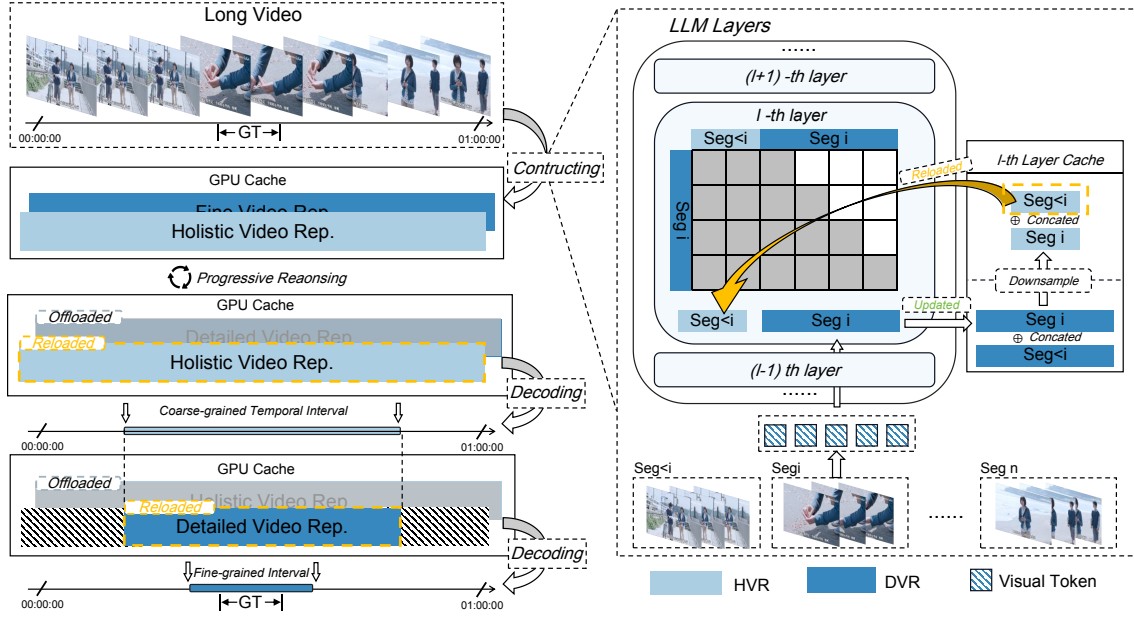


Figure 2: Overview of EviScan. The input long video is processed to generate two representations: the Holistic Video Representation (HVR), which captures global context, and the Fine Video Representation (DVR), which retains detailed local information. EviScan first performs coarse-grained reasoning using HVR to narrow the search space, and then refines the localization using DVR within the identified temporal interval to achieve precise task-oriented localization.

strong frontier MLLMs independently predict candidate temporal intervals. We compute pairwise IoU among the predicted intervals; if their agreement exceeds a predefined threshold, we treat the localization as reliable. We then feed the corresponding local clip (together with the answer) into one expert model to produce a final, precise interval.

(3) Localization rationale generation. Conditioned on the question, answer, full-video context, and the verified interval, a reasoning-capable MLLM generates a chain-of-thought that simulates the localization process, explicitly connecting visual evidence to the task requirement. Finally, we apply an LLM-based rewriting stage to improve the clarity and consistency of the generated rationales, yielding the final localization rationales used for training.

Using this pipeline, we obtain 80K high-quality training samples for EviScan. In addition, we collect and curate 100K off-the-shelf temporal grounding samples. We use all data for supervised fine-tuning (SFT). Additional data and training details are provided in the Appendix.

3.3 DualKV Prefill

Every second matters. In long-video understanding (LVU) tasks, the evidence needed to answer a

question may appear only within a few seconds of an hour-long video. A localization module must therefore *scan densely in time* to avoid missing decisive moments. Yet dense sampling collides with a fixed memory budget: naively increasing the number of frames either exceeds the context window or forces aggressive downsampling that blurs local details. Can we have both—second-level coverage *and* fine-grained evidence? To this end, we introduce **DualKV Prefill**.

Streaming processing with dual-resolution caches. Rather than encoding all frames in a single pass, **DualKV Prefill** processes a densely sampled video stream under a small, fixed memory budget. As illustrated in Figure 2 (right), we partition the frame sequence into temporal segments $\{S_0, S_1, \dots, S_n\}$ and process them sequentially. For each segment S_i and each transformer layer l , we cache its key-value states as a *Detailed Video Representation* (DVR), denoted DVR_i^l , which preserves high-resolution local evidence. We then temporally downsample DVR_i^l to obtain a compact *Holistic Video Representation* (HVR), denoted HVR_i^l , summarizing the segment-level semantics at substantially reduced resolution.

To incorporate long-range context efficiently, we concatenate all previously generated holistic caches $\{HVR_j^l \mid j < i\}$ into a lightweight historical mem-

Table 1: ToMo-Bench results grouped by video duration (short, medium, and long) and task type (Perception and Reasoning).

Method	Short (<3min)		Medium (3–10min)		Long (>10min)		Perception	Reasoning
	R1@0.3	R1@0.5	R1@0.3	R1@0.5	R1@0.3	R1@0.5	R1@0.5	R1@0.5
Temporal Grounding Models								
TimeR1-7B (Wang et al., 2025c)	20.1	16.0	6.8	4.1	15.8	11.0	11.8	10.8
VideoChatR1-7B (Li et al., 2025c)	17.3	14.0	6.6	5.6	12.1	10.9	10.6	9.2
Temporal-RLT-7B (Li et al., 2025a)	54.0	35.3	17.9	12.3	14.1	11.5	18.0	21.7
UniTime (Li et al., 2025d)	52.4	42.7	29.6	23.2	24.2	22.6	32.4	30.6
General MLLMs								
Qwen-2.5VL-7B (Bai et al., 2025)	48.2	40.6	12.2	10.3	14.6	12.3	24.4	22.1
Keye-VL-1.5-8B (Yang et al., 2025)	36.0	34.7	25.4	25.4	32.7	28.6	29.8	37.2
EviScan-7B	52.3	46.3	45.2	42.3	43.2	39.7	45.9	40.0

ory. The current detailed states DVR_i^l attend to this holistic memory via cross-attention, enabling the model to integrate earlier context without attending over all past detailed frames. The updated states are then forwarded to the next layer to produce DVR_i^{l+1} .

Outputs for Progressive Localization. After processing all segments, **DualKV Prefill** outputs two synchronized representations: (i) a compact *HVR* that supports efficient coarse localization over hour-long videos, and (ii) a fine-grained *DVR* that retains second-level details for precise boundary refinement.

3.4 Progressive Localization

Although **DualKV Prefill** enables second-level coverage over the full video, the target moment only exists in the detailed *DVR* and cannot be automatically found. Directly searching within *DVR* is difficult: fine-grained visual tokens dilute attention, making it hard for the model to focus on decisive evidence. To address this, we propose **Progressive Localization**, a coarse-to-fine localization mechanism for long videos.

Coarse localization: We load only the compact holistic *HVR* into GPU memory, keeping the high-resolution *DVR* offloaded. Conditioned on the task query, EviScan predicts a coarse candidate interval, narrowing down the search space.

Fine localization: We offload *HVR* and reload only the corresponding *DVR* segments within the candidate interval to refine the temporal boundaries. This avoids expensive full-video fine-grained inference while preserving the key local evidence required for precise grounding.

In conclusion, Progressive Localization first obtains a coarse candidate from *HVR* and then refines

boundaries using the *DVR* within the candidate, enabling EviScan to localize the target moment both efficiently and accurately.

4 ToMo-Bench

Existing temporal grounding benchmarks mainly focus on explicit-description queries and relatively short videos, which does not fully reflect the challenges of task-oriented localization in realistic long-video scenarios. To bridge this gap, we introduce **ToMo-Bench**, where the model must localize moments implied by task requirements rather than by simple lexical matching. We construct ToMo-Bench under two principles: *high diversity* and *high annotation quality*.

High Diversity: We collect question–answer pairs from public long-video datasets (Zhou et al., 2025; Fu et al., 2025; Wu et al., 2024; Cheng et al., 2025), set minimum quotas by video duration and task type¹, and also remove unsuitable samples (e.g., summarization-style questions without precise temporal evidence).

High Annotation Quality: We adopt a semi-automated but human-verified pipeline. For each selected sample, Gemini-2.5-Pro first proposes multiple candidate intervals at different temperatures. Annotators then manually verify and refine the temporal boundaries; samples that cannot be reliably grounded are discarded. This process improves boundary precision.

Dataset Statistics: ToMo-Bench covers videos ranging from seconds to over one hour (13.5 minutes on average), spans 32 real-world domains, and includes two task types, each with multiple subtasks, for a total of 12 subtask categories.

¹We follow the Video-MME Definition for video duration and task type.

In conclusion, ToMo-Bench provides a more realistic and challenging evaluation protocol for task-oriented temporal grounding, serving as a stronger testbed for future long-video localization models and agentic video understanding systems. See Appendix for more details.

5 Experiment

5.1 Experimental Settings

Benchmarks. To comprehensively evaluate the capabilities of our model, we conducted tests on benchmarks across different categories, including long video understanding, video temporal grounding, and video question answering. For long video understanding, we utilized two benchmarks: MLVU (Zhou et al., 2025) and LongVideoBench (Wu et al., 2024). For temporal grounding, we employed two traditional benchmarks, Charades-STA (Sigurdsson et al., 2016) and ActivityNet (Caba Heilbron et al., 2015), as well as a long video temporal grounding benchmark, V-STaR (Cheng et al., 2025), and ToMo-Bench. To highlight the model’s long video capabilities, we only used segments longer than 5 minutes from V-STaR.

Metric. For the VideoQA task, we use accuracy to evaluate the model. For the temporal grounding task, we employ the Intersection-over-Union (IoU) metric and its variants. These metrics quantify the alignment between the predicted temporal window T^{pred} and the ground truth temporal window T^{gt} . Detailed definitions can be found in the appendix.

Implementations. For all benchmarks, EviScan samples videos at 1 fps, with a maximum of 2000 frames. We obtain HVR by mean-pooling with a temporal downsampling factor of 4. For the retrieval model-augmented video understanding task, we adopt the default configurations for all models, setting the output time window to 64 seconds. If a model’s output time window is less than 64 seconds, we pad it by centering on the model’s output.

5.2 Leading Moment Localization Performance

To validate that EviScan can serve as a strong moment localization module for long-video understanding, we first organize results into two parts: (1) our task-oriented benchmark ToMo-Bench as the primary evidence, and (2) conventional temporal grounding benchmarks as complementary verification.

5.2.1 (1) Primary Results on ToMo-Bench

We evaluate EviScan on ToMo-Bench. We compare EviScan against both specialized temporal grounding models and general MLLMs.

(a) Performance across Video Durations. As shown in Table 1, EviScan demonstrates strong and consistent performance across all video durations. On medium and long videos, EviScan outperforms all baselines by 10–20 points, demonstrating substantial advantages in handling task-oriented temporal grounding in extended contexts. On short videos, EviScan achieves competitive performance. These results highlight EviScan’s effectiveness in task-oriented temporal localization, particularly its ability to perform accurate reasoning and grounding in long-video scenarios.

(b) Performance across Task Types. From a task perspective, EviScan also demonstrates strong performance across all categories. Notably, in Reasoning tasks (e.g., Temporal Reasoning and Action Reasoning), EviScan outperforms Keye-VL-1.5-8B by 2.8 points on R1@0.5 and exceeds all other baselines by at least 10 points, highlighting its superior task-oriented reasoning and temporal localization. On the Perception task, EviScan also achieves state-of-the-art performance.

5.2.2 (2) Complementary Results on Others.

Beyond ToMo-Bench, we also present evaluation results on conventional temporal grounding benchmarks, covering both short-video and long-video settings.

(a) Short-video Benchmarks. As shown in Table 2, EviScan achieves state-of-the-art performance on both Charades-STA and ActivityNet. Notably, compared to most baselines, EviScan maintains a smaller gap between R1@0.5 and R1@0.7, indicating its capability for more precise temporal localization.

(b) Long-video Benchmarks. Moreover, EviScan achieves an R1@0.7 score of 81.8 on V-STaR (with videos duration > 300s), demonstrating EviScan’s strong capability in handling extended temporal grounding tasks.

5.3 Benefits to Long-Video Understanding

To validate the effectiveness of EviScan for long-video understanding, we study whether improved moment localization translates into stronger downstream LVU performance when paired with strong MLLM backbones. We benchmark against a comprehensive set of baselines, including multimodal

Table 2: Performance comparison on short video temporal grounding tasks (including Charades-STA and ActivityNet) and long video temporal grounding (V-STaR).

Method	Charades-STA			ActivityNet			V-STaR(long)	
	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7
Temporal Grounding Models								
TimeR1-7B (Wang et al., 2025c)	72.2	50.1	-	58.6	39.0	48.0	-	-
VideoChatR1-7B (Li et al., 2025c)	71.7	50.2	60.8	33.4	17.7	36.6	-	-
Temporal-RLT-7B (Li et al., 2025a)	67.9	44.1	57.0	38.4	20.2	39.0	-	-
UniTime (Li et al., 2025d)	75.3	56.9	-	54.8	36.6	-	62.9	62.9
General MLLMs								
Qwen-2.5VL-7B (Bai et al., 2025)	-	-	43.6	-	-	-	0.0	0.0
Keye-VL-1.5-8B (Yang et al., 2025)	-	-	-	-	-	-	63.9	49.1
EviScan-7B	75.6	56.0	67.1	59.5	42.7	55.6	86.4	81.8

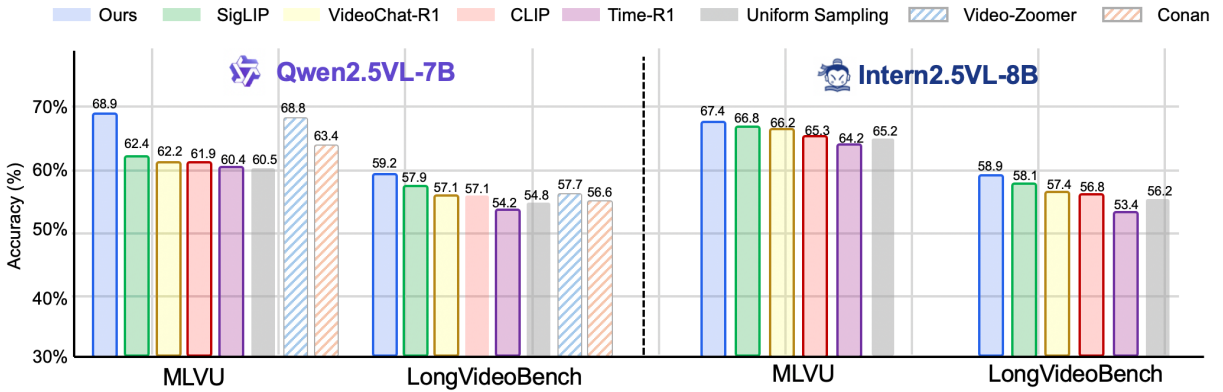


Figure 3: Performance comparison of different VLMs using various keyframe localization methods on MLVU and LongVideoBench.

retrievers, temporal grounding models, and tool-augmented video agents.

Evaluation protocol. For multimodal retrievers, temporal grounding models, and EVISCAN, we first use them to predict a temporal interval. We then sample frames within the predicted interval and feed them into the same MLLM backbone to answer the question. For tool-augmented video agents (e.g., Video-Zoomer and Conan) who solve the task end-to-end with multi-turn tool invocation, we report the performance score from their papers.

As shown in Figure 3, different localization modules yield clearly different gains on long-video benchmarks when applied to Qwen2.5-VL-7B and Intern2.5VL-8B. EVISCAN consistently achieves the largest improvements across backbones and datasets. Notably, EVISCAN remains competitive even against strong multi-turn agentic approaches, highlighting its effectiveness as a standalone moment-localization component.

Overall, these results indicate that EVISCAN is a

strong candidate localization tool for video agents, and motivate future work on integrating it into agentic LVU.

5.4 Ablation Studies

Near-second Temporal Scanning. Figure 4 studies how performance and efficiency scale with the number of input frames. With DualKV Prefill enabled, EviScan can process up to 2000 frames without memory overflow. Increasing input from 800 to 2000 frames improves R@1 at IoU 0.7 from 61.4% to 76.1%, indicating that denser temporal coverage reduces the chance of missing decisive moments. Meanwhile, inference time increases smoothly (e.g., ~ 2.5 s $\rightarrow$ ~ 8.2 s), demonstrating good scalability. In contrast, disabling DualKV Prefill causes memory overflow for long inputs, preventing near-second scanning in hour-long videos.

Effectiveness of Progressive Localization. To evaluate the effectiveness and necessity of progressive localization in long-video scenarios, we con-

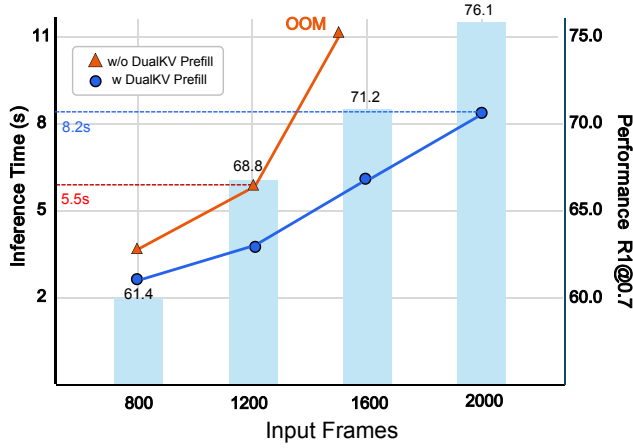


Figure 4: Scaling behavior of EviScan with and without DualKV Prefill.

Table 3: Comparison of progressive localization versus standard prediction on long-video subsets.

Method	V-STaR(long)		ToMo-Bench(long)	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
Zero-shot	73.2	61.4	37.8	29.5
Progressive Localization	86.4	81.8	43.2	39.7

Table 4: Ablation study for the downsampling factor of Holistic Video Representation. D.F. means downsampling factor.

HVR D.F.	V-STaR(long)		ToMo-Bench(long)	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
1	73.2	61.4	37.8	29.5
2	80.7	76.1	41.0	37.1
4	86.4	81.8	43.2	39.7

Table 5: Analysis of training effect from ToMo-Pile. Performance on the Charades-STA, ToMo-Bench (long), and V-STaR bench (duration > 300).

Method	Charades-STA		V-STaR(long)		ToMo-Bench(long)	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7
W/o ToMo-Pile	71.2	49.8	47.7	39.7	29.4	22.7
W ToMo-Pile	75.6	56.0	86.4	81.8	43.2	39.7

ducted an ablation study comparing progressive localization versus standard single prediction. Results are presented in Table 3. The findings demonstrate that progressive localization achieves substantial improvements over the single-step baseline across both traditional temporal grounding and task-oriented tasks in long videos. The performance gain in IoU@0.7 confirms the effectiveness of progressively narrowing the search space for precise temporal localization.

Downsampling Factor for HVR. Table 4 ablates the downsampling factor (D.F.) used to form the holistic video representation (HVR; i.e., holistic KV caches), while keeping the total input budget fixed at 2000 frames and evaluating on long-video subsets. Increasing D.F. consistently improves localization accuracy on both V-STaR(long) and ToMo-Bench(long). We hypothesize this gain comes from two effects: (i) a larger D.F. shortens the holistic sequence, making long-range input easier under positional constraints; and (ii) temporal aggregation suppresses redundant or noisy frames, yielding a cleaner global summary for the coarse-stage reasoning in Progressive Localization.

Effectiveness of ToMo-Pile Dataset. To further analyze the impact of the ToMo-Pile dataset, we

retrained the model by excluding it from the training set and reported the corresponding results on short-video temporal benchmark Charades-STA, long-video benchmarks V-STaR and ToMo-Bench (long), as shown in Table 5. It can be observed that the performance gap becomes particularly pronounced on V-STaR and ToMo-Bench (long). This demonstrates the crucial value of ToMo-Pile in establishing EviScan’s long-video understanding capability and task-oriented reasoning proficiency.

6 Conclusion

In this work, we propose EviScan, an evidence-seeking moment localization model. It leverages DualKV Prefill for memory-efficient streaming inference and incorporates a progressive localization strategy to efficiently perform evidence-seeking moment localization on long videos. To facilitate task-oriented localization learning, we construct ToMo-Pile, a training dataset that pairs temporal intervals with rationale traces explaining why the interval resolves the task. We further introduce ToMo-Bench. Extensive experiments demonstrate that EviScan consistently outperforms specialized localizers and general-purpose MLLMs on both ToMo-Bench and standard temporal ground-

ing benchmarks. We hope this work will advance the deeper understanding of the temporal dimension in videos for MLLMs.

7 Limitation

EviScan has two main limitations. First, when the coarse-stage reasoning fails to correctly identify certain target parts, the subsequent fine-grained representation seems to lack the ability to recover from such errors. One possible solution is to allow the model to fall back to the coarse stage when it determines, during the second stage, that it cannot resolve the task. Second, when the answer to a question is distributed across multiple non-contiguous video segments, the current method can only localize the most representative segment, failing to cover all relevant pieces. This is primarily because constructing multi-segment moment localization data is extremely difficult, and verifying and correcting the annotations is even more challenging. Future efforts may require the development of stronger understanding models as tool models to generate such data. We will work on addressing these issues in future research.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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](#). *Preprint*, arXiv:2502.13923.
- Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2025. Flexible frame selection for efficient video reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29071–29082.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970.
- Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. 2025. [V-star: Benchmarking video-llms on video spatio-temporal reasoning](#). *Preprint*, arXiv:2503.11495.
- Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujie Yang. 2025. [Videozoomer: Reinforcement-learned temporal focusing for long video reasoning](#). *Preprint*, arXiv:2512.22315.
- Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2024. Finevideo.
- 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, Yuhao 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 Computer Vision and Pattern Recognition Conference*, pages 24108–24118.
- Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. 2025a. [Reinforce-gpt: Reinforcing video reasoning in mllms: Reward design and data efficiency](#). *Preprint*, arXiv:2506.01908.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *Preprint*, arXiv:2301.12597.
- Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhao Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025b. [Videochat-flash: Hierarchical compression for long-context video modeling](#). *Preprint*, arXiv:2501.00574.
- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025c. [Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning](#). *Preprint*, arXiv:2504.06958.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. [Llama-vid: An image is worth 2 tokens in large language models](#). *Preprint*, arXiv:2311.17043.
- Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2025d. [Universal video temporal grounding with generative multi-modal large language models](#). *Preprint*, arXiv:2506.18883.
- Zhengyang Liang, Yan Shu, Xiangrui Liu, Minghao Qin, Kaixin Liang, Nicu Sebe, Zheng Liu, and Lizi Liao. 2025. Video-browser: Towards agentic open-web video browsing. *arXiv preprint arXiv:2512.23044*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025. [Video-xl-pro: Reconstructive token compression for extremely long video understanding](#). *Preprint*, arXiv:2503.18478.

- Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. [Conan: Progressive learning to reason like a detective over multi-scale visual evidence](#). *Preprint*, arXiv:2510.20470.
- Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. 2025. [Video-xl-2: Towards very long-video understanding through task-aware kv sparsification](#). *Preprint*, arXiv:2506.19225.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multi-modal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.
- Xiaoqian Shen, Yuniang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. [Longvu: Spatiotemporal adaptive compression for long video-language understanding](#). *Preprint*, arXiv:2410.17434.
- Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2024. [Video-xl: Extra-long vision language model for hour-scale video understanding](#). *Preprint*, arXiv:2409.14485.
- Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. [Adaptive keyframe sampling for long video understanding](#). *Preprint*, arXiv:2502.21271.
- Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. 2025a. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. [Videoagent: Long-form video understanding with large language model as agent](#). *Preprint*, arXiv:2403.10517.
- Xidong Wang, Dingjie Song, Shunian Chen, Junyin Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. 2025b. [Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture](#). *Preprint*, arXiv:2409.02889.
- Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. 2025c. [Time-r1: Post-training large vision language model for temporal video grounding](#). *Preprint*, arXiv:2503.13377.
- 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.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.
- Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, and 42 others. 2025. [Kwai keye-vl 1.5 technical report](#). *Preprint*, arXiv:2509.01563.
- Zhoufaran Yang, Yan Shu, Jing Wang, Zhifei Yang, Yan Zhang, Huaying Yuan, Yu Li, Keyang Lu, Gangyan Zeng, Shaohui Liu, and 1 others. 2026. Vidtext: Towards comprehensive evaluation for video text understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7575–7585.
- Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. 2025. [Generative frame sampler for long video understanding](#). *Preprint*, arXiv:2503.09146.
- Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. 2025. [Frame-voyager: Learning to query frames for video large language models](#). *Preprint*, arXiv:2410.03226.
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. [Long context transfer from language to vision](#). *Preprint*, arXiv:2406.16852.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. [Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment](#). *Preprint*, arXiv:2310.01852.

Overview of Appendix

- [A: Different and Other Grounding-QA](#)
- [B: ToMo-Pile detail](#)
- [C: ToMo-Bench detail](#)
- [D: Metric](#)
- [E: Experimental Settings & Additional Results](#)
- [F: VideoQA Task Details](#)
- [G: Limitations & Future Work](#)
- [H: Broader Impacts Statement](#)
- [I: Qualitative Results](#)
- [J: Prompt for ToMo-Pile Generation](#)

A Different and Other Grounding-QA

Although some traditional groundingQA datasets—such as CG-Bench, NExT-GQA, and ReXTime already require localizing temporal evidence based on the question rather than on explicit event phrases, the Task-oriented Temporal Grounding task we propose is fundamentally different. These differences are illustrated in Figure 1. Specifically, questions in conventional question-based grounding works usually still contain words strongly correlated with the target time interval, allowing existing temporal-grounding models to localize the interval by relying on the dominant cues in the question. In contrast, our task-oriented setting emphasizes reasoning and thinking driven by the question: our questions typically lack descriptors that are directly tied to the answer, and instead demand that the model first infer which temporal segments can solve the question and then ground them.

B ToMo-Pile detail

Our ToMo-Pile comprises 80k high-quality COT grounding data and 100k long video grounding data, where the COT data originates from VideoR1 and FineVideo, re-annotated using the method described in the paper, and the long video grounding data is constructed by concatenating video clips from pandas-70M.

Figure 2 presents the statistics for ToMo-Pile. The 80k CoT data is sourced from VideoR1 and

FineVideo, with their video duration and word count shown in the figure. Additionally, the 100k long-video grounding data all have a duration of 600 seconds.

C ToMo-Bench Detail

ToMo-Bench comprises 337 videos and nearly 500 questions. As shown in Figure 3, it exhibits rich diversity in task types, video categories, video durations, and the temporal locations of target intervals. The average video length is 805 seconds. Specifically, the benchmark covers 12 task types—such as action reasoning, OCR perception, and temporal reasoning—spanning a wide range of video-understanding tasks.² In addition, it includes 32 video categories from numerous real-world domains. Video durations range from a few minutes to one hour, and the target intervals are uniformly distributed along the timeline to ensure an unbiased evaluation.

Figure 4 shows the word cloud of ToMo-Bench. **Explanations of the 12 Task Categories in ToMo-Bench:**

C.1 Reasoning

- **OR (Object Reasoning):** Identify why an object is present and how it causally affects later events in the clip.
- **SR (Spatial Reasoning):** Infer spatial relationships (e.g., behind, left-of) between objects that are never simultaneously visible.
- **TR (Temporal Reasoning):** Determine the correct chronological order of events shown out of sequence or with missing segments.
- **AR (Action Reasoning):** Predict the next action an agent will perform given the current context and goals.

C.2 Perception

- **SP (Spatial Perception):** Locate and describe where objects are in the 3-D scene from ego-centric or exocentric views.
- **AP (Attribute Perception):** Recognize object properties such as color, material, or state that remain constant across frames.

²We adopt the definitions of task type and video category from Video-MME.

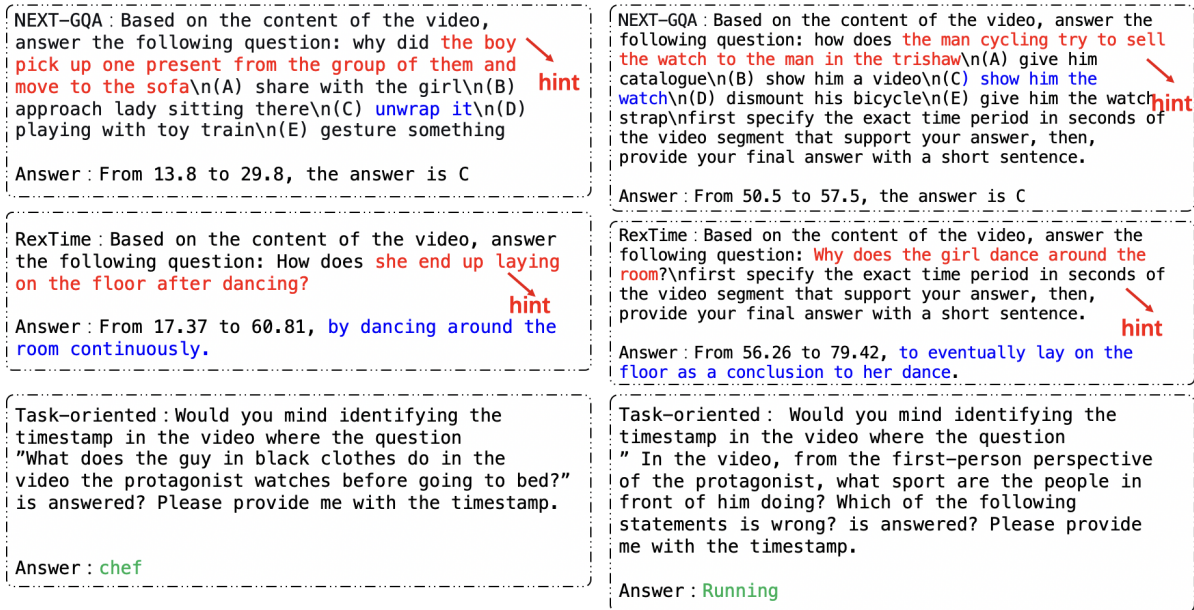


Figure 1: Different and Other Grounding-QA. Red denotes the hint to the answer span in the question, whereas blue marks the corresponding part in the answer. Task-oriented questions contain no cue about the target span, while all other grounding-QA tasks do.



Figure 2: Statistics for ToMo-Pile, including video duration, word cloud, and word count.

- **TP (Temporal Perception):** Detect when a change or event occurs and report its exact start and end times.
- **EG (Ego):** Understand what the camera-wearer is doing, attending to, or will do next from first-person video.
- **OP (OCR Problems):** Read and transcribe text that appears on screens, signs, or documents within the video.
- **AT (Action Recognition):** Classify which predefined action is being performed in a given segment of the video.
- **CP (Counting Problem):** Accurately count how many instances of a specified object or person appear throughout the clip.
- **TU (Tutorial):** Follow instructional video content and answer questions about the steps

or techniques demonstrated.

C.3 Duration

For length classification, we adopt a three-tiered approach: videos shorter than 3 minutes are classified as 'short', those between 3-10 minutes as 'medium', and those longer than 10 minutes as 'long'.

D Metric

To evaluate the performance of our model on temporal grounding and VideoQA tasks, we employ the Intersection-over-Union (IoU) metric and its variants. These metrics quantify the alignment between the predicted temporal window T^{pred} and the ground truth temporal window T^{gt} . The formal definitions are as follows:

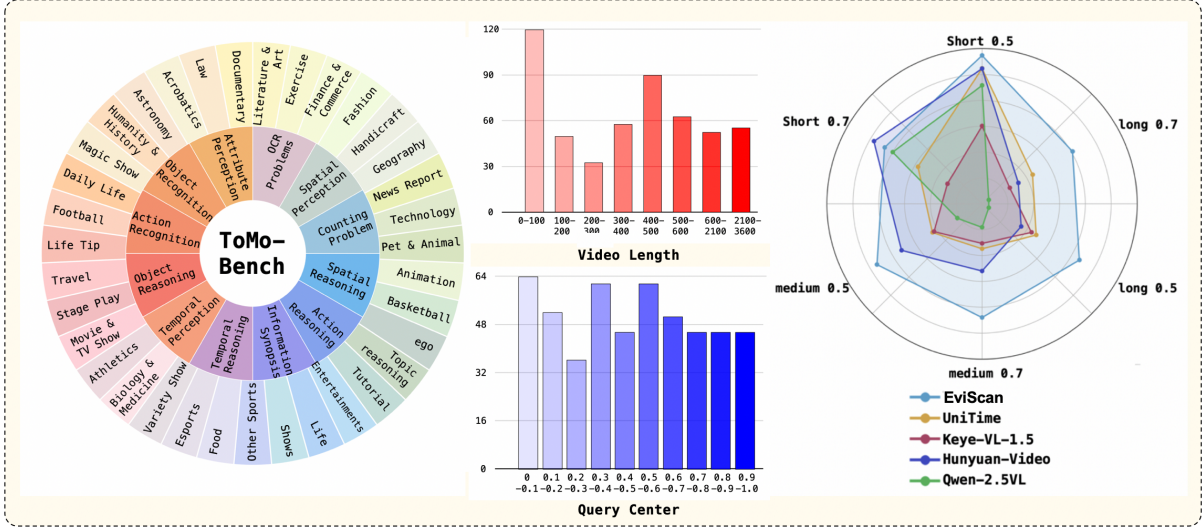


Figure 3: Statistics analysis of ToMo-Bench. (Left) Our benchmark covers distinct task types and 32 video categories. (Middle) Video duration and question center distributions. (Right) Performance of various models on ToMo-Bench.



Figure 4: Wordcloud of ToMo-Bench, left is Question, right is answer.

D.1 IoU and mIoU

The fundamental IoU metric is defined as:

$$\text{IoU} = \frac{|T^{\text{pred}} \cap T^{\text{gt}}|}{|T^{\text{pred}} \cup T^{\text{gt}}|}$$

The mean Intersection-over-Union (mIoU) is calculated as the average IoU across all test samples:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \text{IoU}_i$$

where N is the total number of test samples and IoU_i is the IoU value for the i -th sample.

We also evaluate performance using IoU thresholds, which measure the percentage of predictions exceeding specific IoU values:

- **IoU@0.3:** Percentage of predictions with $\text{IoU} \geq 0.3$

$$\text{IoU@0.3} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{IoU}_i \geq 0.3) \times 100\%$$

- **IoU@0.5:** Percentage of predictions with $\text{IoU} \geq 0.5$

$$\text{IoU@0.5} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{IoU}_i \geq 0.5) \times 100\%$$

- **IoU@0.7:** Percentage of predictions with $\text{IoU} \geq 0.7$

$$\text{IoU@0.7} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{IoU}_i \geq 0.7) \times 100\%$$

Here, $|T^{\text{pred}} \cap T^{\text{gt}}|$ denotes the duration of the overlapping region between the predicted window T^{pred} and the ground truth window T^{gt} . $|T^{\text{pred}} \cup T^{\text{gt}}|$ represents the duration of their union, while $|T^{\text{pred}}|$ and $|T^{\text{gt}}|$ denote the durations of the predicted and ground truth windows, respectively. The indicator function $\mathbb{I}(\cdot)$ equals 1 when the condition is true and 0 otherwise.

Query Center		0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1	diff
Qwen2.5-VL	R@0.3	0.417	0.368	0.615	0.222	0.250	0.158	0.278	0.235	0.133	0.154	78%
	R@0.5	0.250	0.158	0.462	0.185	0.062	0.158	0.111	0.011	0.066	0.000	97%
	R@0.7	0.208	0.158	0.308	0.074	0.000	0.053	0.055	0.000	0.000	0.000	100%
Keye-1.5-VL	R@0.3	0.615	0.652	0.666	0.444	0.375	0.353	0.556	0.600	0.235	0.461	64%
	R@0.5	0.538	0.521	0.400	0.379	0.186	0.294	0.444	0.466	0.176	0.307	67.2%
	R@0.7	0.462	0.260	0.266	0.259	0.125	0.176	0.222	0.333	0.117	0.231	74.6%
Unitime	R@0.3	0.669	0.465	0.633	0.364	0.371	0.216	0.550	0.567	0.371	0.567	67.7%
	R@0.5	0.592	0.335	0.633	0.221	0.294	0.216	0.500	0.566	0.311	0.433	65.0%
	R@0.7	0.476	0.335	0.367	0.186	0.176	0.163	0.500	0.455	0.194	0.433	67.4%
EviScan	R@0.3	0.393	0.417	0.606	0.464	0.421	0.470	0.555	0.667	0.470	0.467	35.1%
	R@0.5	0.393	0.417	0.606	0.357	0.368	0.353	0.505	0.556	0.412	0.400	41.7%
	R@0.7	0.357	0.375	0.533	0.286	0.263	0.294	0.400	0.500	0.353	0.267	50.6%

Hyperparameter	Stage 1	Stage 2
Overall batch size	64	64
Learning rate	1e-5	1e-5
LR Scheduler	Cosine decay	Cosine decay
DeepSpeed ZeRO Stage	ZeRO-2-offload	ZeRO-2-offload
Optimizer	Adam	Adam
Warmup ratio	0.3	0.3
Epoch	1	1
Weight decay	0	0
Precision	bf16	bf16

Table 1: Hyperparameters of EviScan for different training stages

IoU measures the overall alignment between T^{pred} and T^{gt} , providing a balanced assessment of both precision and recall by considering the overlap relative to their union. The threshold-based metrics (IoU@0.3, IoU@0.5, IoU@0.7) evaluate the model’s ability to produce high-quality predictions meeting different precision standards, while mIoU provides an overall average performance measure across all samples.

we adopt the “R1@m” evaluation protocol to compare with state-of-the-art models, which computes the percentage of samples where the top-1 predicted segment has an IoU greater than a threshold m , with $m \in \{0.3, 0.5, 0.7\}$. For brevity, we also adopt mIoU, which calculates the average IoU on all testing data as an alternative metric.

D.2 Query Center Robustness

To measure the effectiveness of balanced query centers in the ToMo-Bench, we evaluated the sensitivity of state-of-the-art (SOTA) models to query

Dataset	Context Length
Charades-STA	1fps
Activity-Net	1fps
Vid-Chapters	1fps(<800 frames)
CG-Bench	1fps(<800 frames)
MLVU	1fps(<800 frames)
LongVideoBench	1fps(<800 frames)
V-STaR	1fps(<2000 frames)

Table 2: Experimental settings of EviScan.

centers. As shown in the figure, experiments indicate that most models perform better on test data with query centers positioned earlier rather than later or in the middle. We measured the difference between the best and worst performance of various models when the query center varies. The results show that EviScan maintains its effectiveness regardless of the position of the query center (for instance, the gap is only 35% at iou@0.3, which is significantly better than Qwen2.5-VL’s 78% and 23.8% higher than UniTime). This further proves the effectiveness of EviScan’s training.

E Experimental Settings & Additional Results

We elaborate on the training and inference details of EviScan. The reported hyperparameters cover Stage 1, and 2, as specified in Table 1. For the inference details, we emphasize the particular context length for different benchmarks, as shown in Table 2.

F VideoQA Task Details

F.1 VideoQA with Temporal Grounding

We use Qwen2.5-VL-7B as the VideoQA model for answer generation. By default, it processes long videos by uniformly sampling 32 frames. However, this sampling strategy may lead to the omission of critical information. To investigate whether temporal grounding models can compensate for this issue, we adopt the following procedure. First, we use different video temporal grounding models to localize the relevant segments for each question. Then, we crop the localized video intervals and input them into Qwen2.5-VL-7B. Specifically, for cropped video segments shorter than 32 seconds, we extend their duration from the center to 32 seconds. Within each interval, we again uniformly sample 32 frames for answer generation.

F.2 Prompt template for VideoQA

We use the same prompt template for all multiple-choice VideoQA benchmarks:

System:

You are a helpful assistant.

User:

<video>

Question: <question>

Options:

(A) <Option_A>

(B) <Option_B>

(C) <Option_C>

(D) <Option_D>

Please only give the best option.

Best Option:

Assistant:

G Limitations & Future Work

Although EviScan demonstrates exceptional performance on various video temporal grounding and video QA benchmarks, it still has several limitations that warrant further exploration: (i) EviScan is currently constrained to temporal grounding tasks (including traditional temporal grounding tasks and Task-Oriented Moment localization tasks). To enable broader applications in MLLMs, it requires more diverse training data with dense temporal annotations. Incorporating such data into the pretraining process of MLLMs could unlock their potential for handling more temporally complex tasks, such as dense video captioning. (ii) Although EviScan enhances MLLMs with temporal

grounding capabilities, relying solely on temporal grounding data limits their reasoning and question-answering abilities. The ultimate objective is to develop MLLMs that seamlessly integrate localization, reasoning, and question-answering into a unified framework.

H Broader Impacts Statement

We have developed the ToMo-Bench evaluation suite, designed to assess Moment localization performance across diverse real-world long video understanding scenarios, thereby fostering progress in this emerging field. This initiative will enhance the complexity of video temporal understanding and accelerate the development of models that integrate reasoning capabilities with traditional temporal grounding functions. Our ultimate goal is to unify reasoning and temporal grounding within a single framework. Concurrently, we have introduced EviScan, a more accurate and efficient Moment localization framework, to advance research in long video temporal understanding. The outcomes of this work are expected to benefit a wide range of downstream applications, including anomaly detection and security surveillance. We firmly believe that ToMo-Bench and EviScan will jointly propel the advancement of video temporal understanding technologies.

I Qualitative Results

In Figure 5–9, we present the qualitative results of EviScan on ToMo-Bench and V-STaR. EviScan demonstrates accurate understanding of questions and the ability to provide temporal grounding. In V-STaR, we show the results when two-stage reasoning is applied, and it can be seen that EviScan exhibits robust performance with good coarse-grained segment retrieval and fine-grained temporal grounding capabilities.

J Prompt for ToMo-Pile Generation

In Figure 10–12, shows the prompts we used to construct ToMo-Pile with all expert models. Figure 10 illustrates the prompt for filtering tasks using Qwen. Figure 11 presents the prompts for grounding using Keye-1.5VL and ARC-Hunyuan-Video. Figure 12 displays the prompt for Qwen to summarize and rewrite the chain of thought generated by Keye-1.5VL.

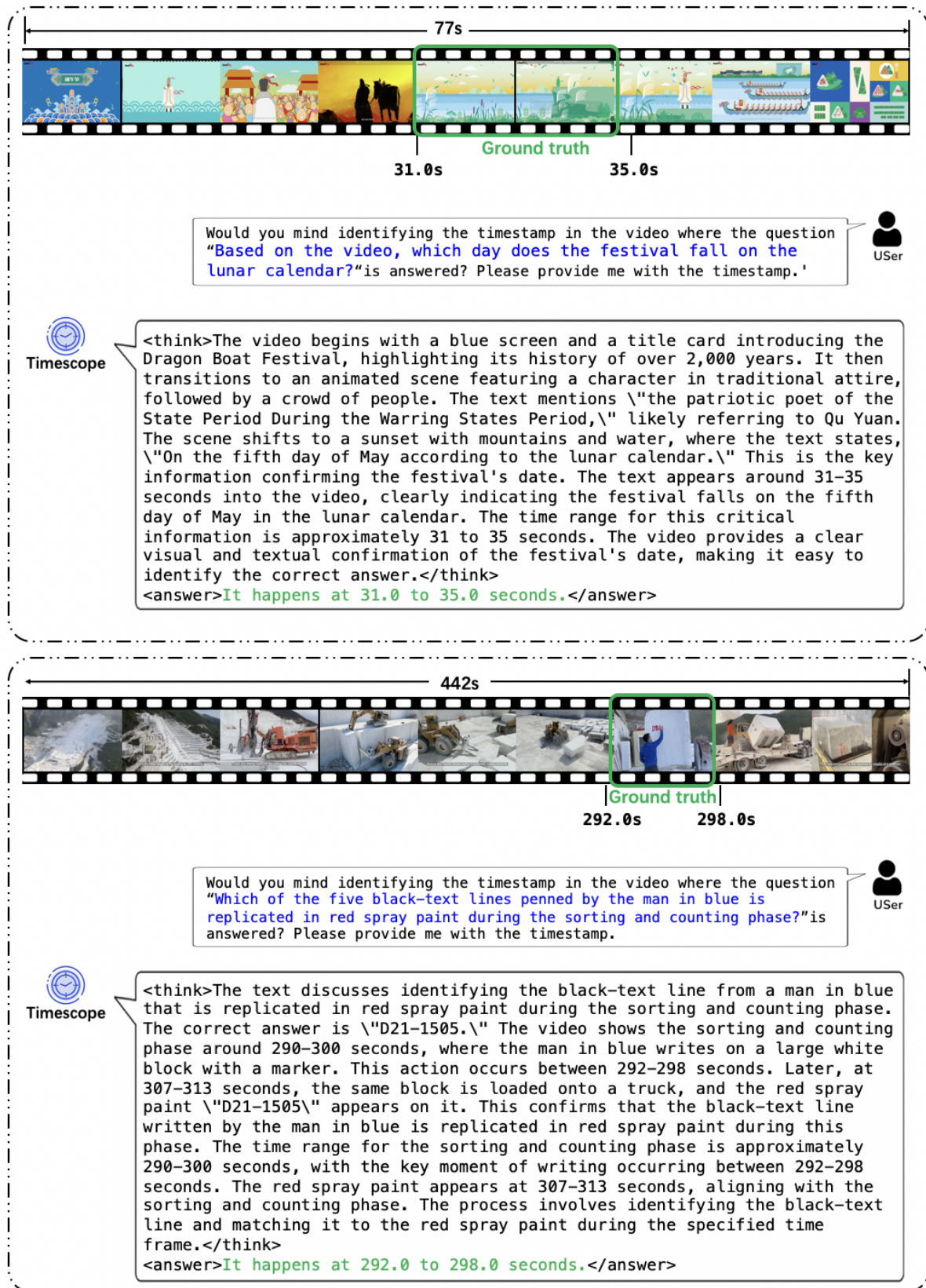


Figure 5: Qualitative Results of EviScan.

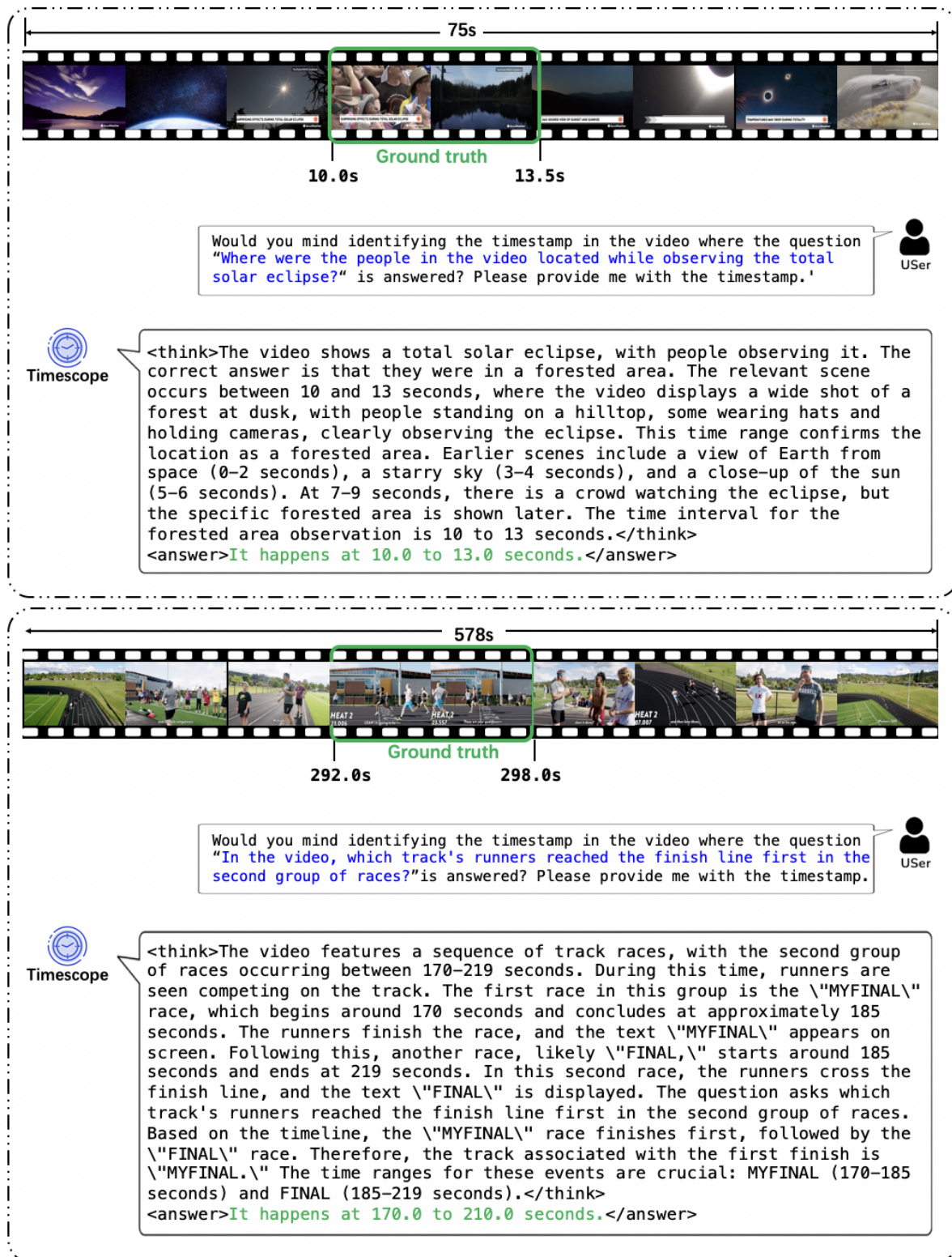


Figure 6: Qualitative Results of EviScan.



Figure 7: Qualitative Results of EviScan.



Figure 8: Qualitative Results of EviScan.

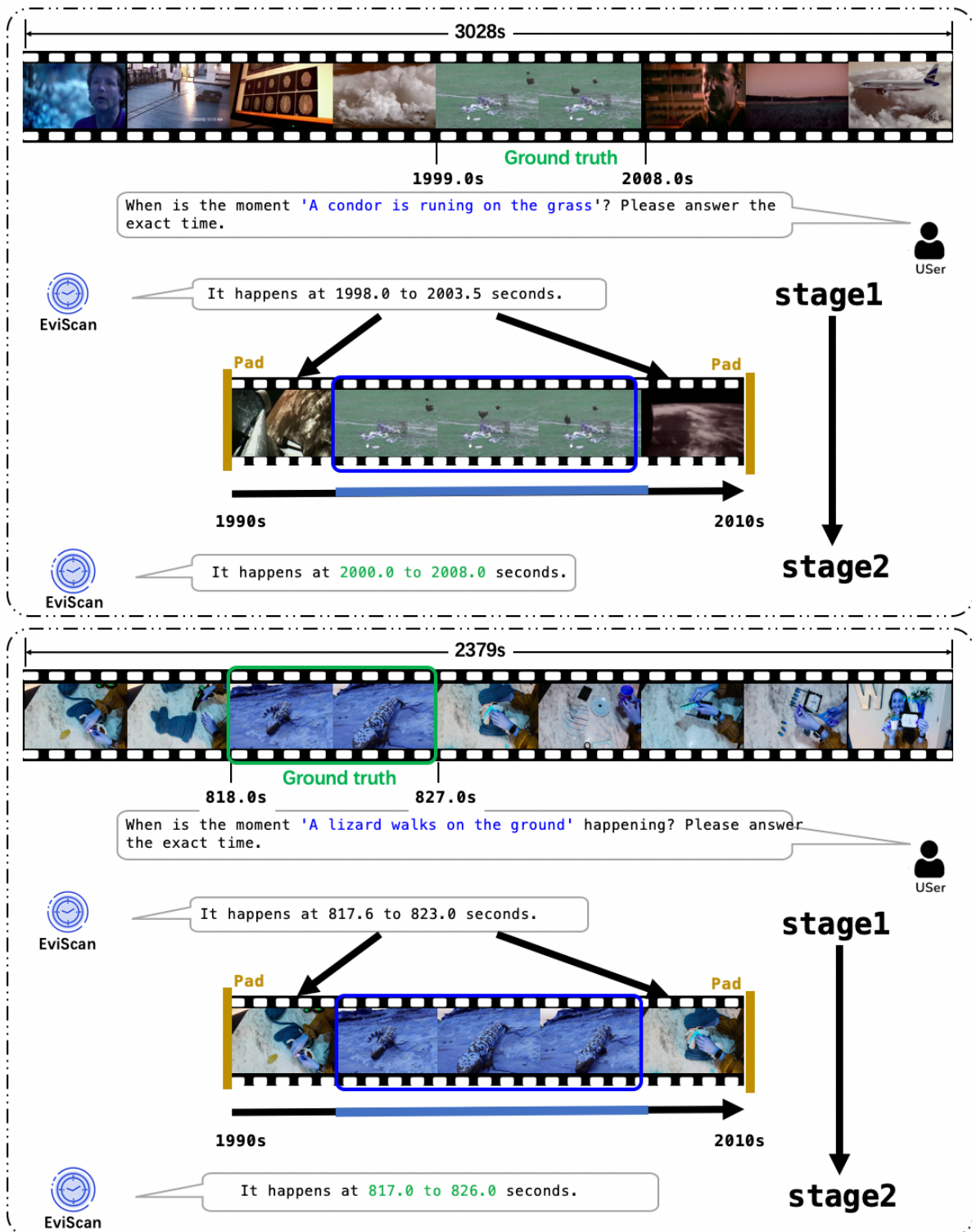


Figure 9: Qualitative Results of EviScan.

System Prompt for Task Filter

For a given video question, you need to determine whether it can be answered based solely on a single segment of the video, rather than requiring global information from the entire video. Such questions typically belong to detailed question-answering tasks, such as asking about an action at a specific moment, the color of an object, or the content of a particular conversation. In contrast, global summarization tasks require summarizing the video's main idea or describing its overall atmosphere.

If the judgment is "yes", please output "yes"; if the judgment is "no", please output "no." Ensure that the output contains only "yes" or "no," without any additional explanations, punctuation, or spaces.

Question: "question"

Answer: "answer"

Figure 10: System Prompt for Task Filter.

System Prompt for Grounding Expert

You need to complete a video temporal grounding task. I will provide you with content that includes a question and the corresponding correct answer, formatted as "Question: Answer." Your task is to analyze the video content, reconstruct the video timeline, and accurately mark the time ranges (in seconds) for the events mentioned in the correct answer. You should carefully reason through the answer within the identified time range to ensure that the temporal information is as precise as possible. If you are uncertain about any part of the answer, please output "Uncertain."

Question: "question"

Answer: "answer"

Figure 11: System prompt template for Grounding Expert.

System Prompt for LLM Summary of Chain of Thought

Please provide a concise summary of the provided text and return only the summarized version. Your summary must adhere to the following requirements: Focus on extracting and refining the essential information from the text. A critical requirement is to emphasize and retain all time-related details (for example, specific events that occur within a timeframe like "5-10 seconds"). These temporal elements must not be omitted. However, if there is any redundant or repeated information, you may consolidate it by keeping it only once. You are also permitted to appropriately trim or omit any introductory phrases or filler words at the beginning of the text to enhance conciseness. The final summarized paragraph should be strictly controlled to a maximum of 200 words. Please return only the final summarized text without any additional comments or explanations.

Paragraphs: "paragraph"

Figure 12: System Prompt for LLM Summary of Chain-of-Thought.