TimeScope: Towards Task-Oriented Temporal Grounding In Long Videos

Xiangrui Liu[1,2†] Minghao Qin[1†] Yan Shu[3†] Zhengyang Liang[4] Yang Tian[2]
Chen Jason Zhang[1] Bo Zhao[2] Zheng Liu[1]

[1]Beijing Academy of Artificial Intelligence [2]School of AI, Shanghai Jiao Tong University
[3]University of Trento [4]Singapore Management University

†Equal contribution.

Abstract

Identifying key temporal intervals within long videos, known as temporal grounding (TG), is important to video understanding and reasoning tasks. In this paper, we introduce a new form of the temporal grounding problem, ***Task-oriented Temporal Grounding (ToTG)****, which is driven by the requirements of downstream tasks rather than explicit time-interval descriptions. For example, a ToTG input may be "explain why the man in the video is sent to the hospital," whereas traditional TG would take an explicit temporal description such as "the moments when the man is tripped by a stone and falls to the ground." This new ToTG formulation presents significant challenges for existing TG methods, as it requires jointly performing deep task comprehension and fine-grained temporal localization within long videos. To address these challenges, we conduct a systematic set of studies. First, we construct* ***a new benchmark ToTG-Bench****, which comprehensively evaluates ToTG performance across diverse settings. Second, we introduce* ***a new temporal-ground method TimeScope****, which performs coarse-to-fine localization through a progressive reasoning process. Leveraging extensive supervised fine-tuning with carefully curated chain-of-thought (CoT) data from a variety of scenarios, TimeScope generalizes effectively across tasks and domains. Our evaluation demonstrates* ***TimeScope's empirical advantages*** *over existing baselines from three perspectives: (1) substantial improvements in grounding precision, (2) significant benefits to downstream tasks, and (3) strong generalizability across different scenarios. All models, datasets, and source code will be fully open-sourced to support future research in this area.*

1. Introduction

Multimodal large language models (MLLMs) have become increasingly prominent in tackling long-video understanding (LVU) problems. However, they still struggle with complex tasks which call for fine-grained details, especially those sparsely distributed across long videos. One promising strategy to mitigate this problem is to present MLLMs only with crucial temporal intervals which contains relevant information to their LVU tasks. However, existing temporal grounding (TG) methods are primarily designed for tasks with explicit descriptions of time-intervals, such as "*the moments when the man is tripped by a stone and falls to the ground*", instead of directly handling native task requirements, e.g., "*explain why the man in the video is sent to the hospital*". This gap prevents the direct utilization of existing TG methods for LVU tasks, resulting in a severe limitation for real-world applications.

To formalize the above challenge, we define a new problem: Task-oriented Temporal Grounding (**ToTG**), where a model needs to localize crucial temporal intervals that are relevant to a specific downstream task based on its native requirement. For example, as illustrated in Figure 3, given the query "*tell me what happens after cutting the vegetables*," the model identifies the interval corresponding to "*stir-frying the vegetables*." Unlike traditional TG problems, ToTG is more closely aligned with practical applications, as it can be directly conducted to support real-world LVU tasks. The new ToTG problem introduces unprecedented technical challenges for existing methods on TG, as it requires both *in-depth task comprehension* and *fine-grained temporal localization* over long videos.

In this paper, we present a systematic study of the ToTG problem. First, to address the absence of suitable resources for evaluating ToTG performance, we introduce a new benchmark, **ToTG-Bench**. ToTG-Bench incorporates 32 diverse video domains and 12 task categories, with video durations ranging from a few seconds to over an hour. Each instance is constructed through a human–machine collaborative annotation pipeline that ensures high-quality temporal localization. Together, these designs enable ToTG-Bench to provide a comprehensive evaluation of ToTG per-

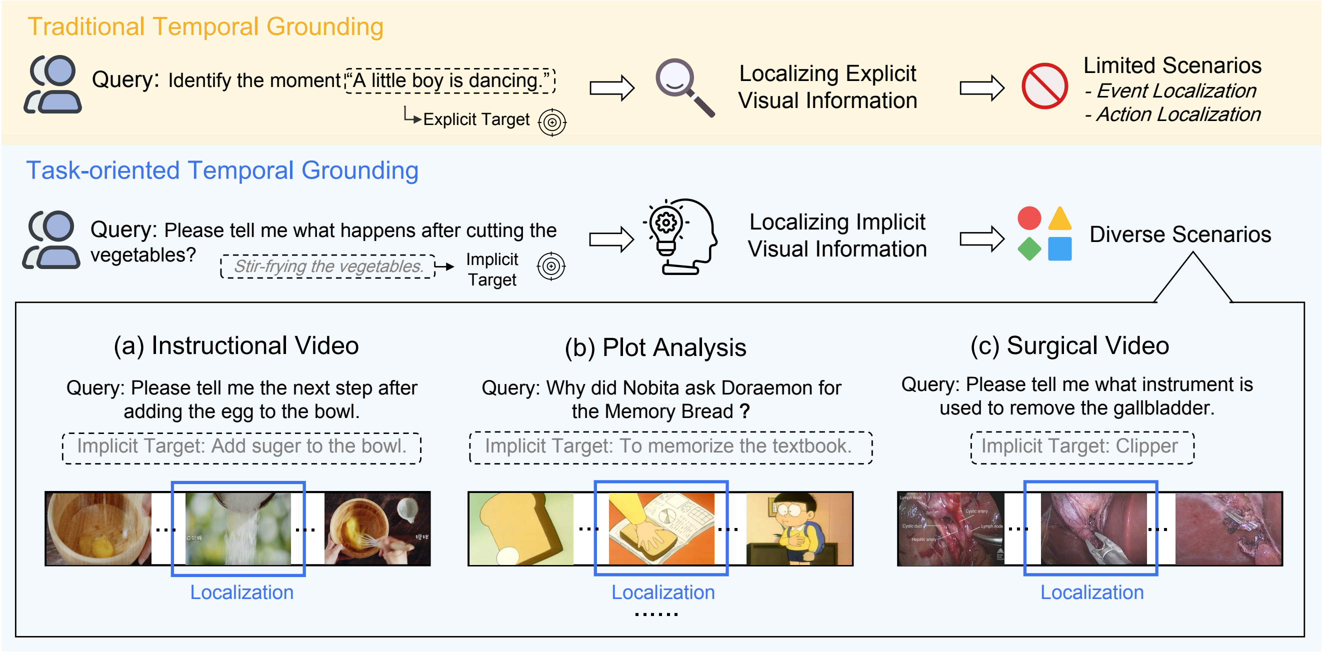

Traditional Temporal Grounding
Query: Identify the moment "A little boy is dancing."
Explicit Target
Localizing Explicit Visual Information
Limited Scenarios
- Event Localization
- Action Localization
Task-oriented Temporal Grounding
Query: Please tell me what happens after cutting the vegetables?
Stir-frying the vegetables.
Implicit Target
Localizing Implicit Visual Information
Diverse Scenarios
(a) Instructional Video
Query: Please tell me the next step after adding the egg to the bowl.
Implicit Target: Add suger to the bowl.
Localization
(b) Plot Analysis
Query: Why did Nobita ask Doraemon for the Memory Bread ?
Implicit Target: To memorize the textbook.
Localization
(c) Surgical Video
Query: Please tell me what instrument is used to remove the gallbladder.
Implicit Target: Clipper
Localization
......

Figure 1. In traditional temporal grounding, the target is explicit and can be located via simple semantic matching, whereas task-oriented temporal grounding requires identifying an implicit target essential for completing the task.

formance across a wide spectrum of task and video types.

To further enhance the grounding model's capability for ToTG, we develop a novel progressive-reasoning framework called **TimeScope**. As discussed, ToTG poses two major technical challenges: 1) in-depth understanding of the task, and 2) fine-grained temporal grounding within long videos. TimeScope addresses these challenges through two consecutive operations. First, the model focuses on comprehending the task and identifying which parts of the video are likely to contain the required information. To accomplish this, TimeScope performs chain-of-thought (CoT) reasoning based on the task description and a holistic abstraction of the video, producing a set of candidate time scopes that are likely to be relevant. Next, TimeScope further encodes the candidate scopes with more detailed visual information, and predicts the precise temporal intervals contained within them. This progressive reasoning pipeline not only improves the grounding precision but also enables more efficient processing of long videos.

To obtain a general temporal-grounding capability across diverse long-video tasks, TimeScope is trained through extensive supervised fine-tuning. To construct the training data, we prompt a teacher model to generate chain-of-thought (CoT) reasoning for a wide range of LVU tasks. We then filter the generated CoT trajectories, retaining only those whose reasoning leads to precise temporal grounding. This process yields **ToTG-Pile**, a diverse and high-quality dataset tailored for ToTG.

In our experiments, we evaluate TimeScope against traditional TG methods [1, 41] using not only the proposed ToTG-Bench, but also standard temporal-grounding benchmarks [2, 6, 33] and long-video understanding benchmarks [39, 47]. Our evaluation demonstrates the effectiveness of TimeScope from three perspectives. First, it substantially improves temporal-grounding precision over existing TG baselines. Second, it exhibits strong generalizability, achieving consistent performance gains across heterogeneous benchmarks. Third, it significantly enhances the performance of downstream LVU tasks when used as a pre-localization module. Extended ablation studies further highlight the contribution of each component, reflecting the validity of our technical design. All resources, including the model, benchmark, dataset, and source code, will be publicly released to facilitate future research.

The contributions of this paper are summarized as follows. 1) We formulate the **ToTG problem**, which formally defines the localization of task-relevant information in long videos. 2) We introduce **ToTG-Bench**, which enables unified and comprehensive evaluation of ToTG performance across diverse tasks and video domains. 3) We propose **TimeScope**, a progressive reasoning framework for ToTG, and construct **ToTG-Pile**, a high-quality supervised fine-tuning dataset that enhances TimeScope's capability. 4) We conduct extensive experiments, demonstrating TimeScope's **empirical effectiveness** in ToTG precision, generalization across benchmarks, and benefits to downstream LVU tasks.

Benchmark	Video Num.	Avg. Duration	Duration Range	Video Domain	Query Type
Traditional Temporal Grounding Benchmark					
Charades-STA	1,331	29.9 s	7.2 s - 1.2 min	Daily Activities	Explicit Description
ActivityNet	4,885	122.0 s	2.3 s - 12.4 s	Daily Activities	Explicit Description
V-STaR	732	1.8 min	15.0 s - 59.2 min	9 Domains	Explicit Description
Clue-grounded QA Benchmark					
CG-Bench	1,219	28.7 min	9.1 min - 1.8 hr	14 Domains	Perception Task
Next-GQA (test)	990	38.7 s	10.0 s - 2.5 min	Daily Activities	Perception Task
Task-oriented Temporal Grounding Benchmark					
ToTG-Bench	337	13.5 min	30 s - 1.2 hr	35 Domains	Perception/Reasoning Task

Table 1. Comparison of ToTG-Bench with previous temporal grounding and clue-grounded QA benchmarks. ToTG-Bench demonstrates superior diversity and comprehensiveness in video characteristics and query types.

2. Related work

2.1. Long Video Understanding

The field of long video understanding (LVU) has developed rapidly in recent years, with many powerful MLLMs emerging, such as VideoChatFlash [20], Video-XL-2 [29], Eagle2.5 [4], and InternVL3 [49]. These models demonstrate strong general video understanding capabilities and serve as versatile backbones for various video tasks. However, precisely capturing fine-grained details within second-level intervals remains a major challenge for current LVU models. To address this, some works introduce additional modules to assist LVU models by identifying key frames [15, 30, 36, 43]. These modules are typically similarity-based and thus lack deeper semantic understanding of the video content, limiting their compatibility with diverse downstream tasks in long video scenarios. In contrast, we take a different approach. We post-train LVU MLLMs on our diverse and high-quality task-oriented grounding dataset, and further implement **TimeScope**, a novel framework designed for progressive task-oriented grounding. This enables the model to efficiently and accurately localize critical time intervals in long videos for a wide range of tasks.

2.2. Video Temporal Grounding

The traditional temporal grounding (TG) task requires models to localize a time interval in a video given a query that explicitly describes the target content. Early approaches are mainly dual-encoder-based, where video and language features are extracted using different pre-trained encoders (e.g., BERT [7], CLIP [31], SigLip [45]), and then fused for time interval xprediction [11, 17, 25–27, 34, 35]. These models lack generalizability and can only be evaluated under few-shot setting across different benchmarks. More recently, researchers have explored using MLLMs for more general temporal grounding [12, 14, 32, 38, 44]. For instance, TimeChat [32] introduces a time-aware frame encoder that binds visual tokens with their corresponding timestamps at the frame level for temporal grounding. Similarly, TimeSuite [44] proposes temporal-adaptive position encoding to strengthen temporal awareness in video representations. Trace [12] designs a specialized encoder and head for timestamp input, while Time-R1 [38] employs a reasoning-guided post-training framework with reinforcement learning and verifiable rewards to improve grounding accuracy. In addition to these specialized MLLMs, recent generic MLLMs (e.g., Qwen2.5-VL [1], Keye-VL-1.5 [42]) have also demonstrated certain capabilities for temporal grounding. However, localizing small intervals in long videos is challenging due to their limited context window. UniTime [22] tackles this by adjusting the frame sampling rate for multi-stage grounding, but its sparse frame sampling can disrupt continuous event semantics, compromising precise temporal grounding. In addition, previous approaches tend to fall short when it comes to more complex and practical grounding tasks. Motivated by these limitations, we introduce the new problem of **task-oriented temporal grounding**, along with a benchmark, a dataset, and a dedicated framework to address it.

3. Problem Definition

We formally define the Task-oriented Temporal Grounding (ToTG) problem in this section. Formally, the input consists of a long video $V = \{f_t\}_{t=1}^{T}$ and a task-oriented natural language query Q_{task}. ToTG assumes that the critical visual information required to answer the task is not explicitly described in Q_{task}. Therefore, the model must localize the temporal interval containing this task-critical but implicit evidence *without relying on direct semantic matches to the query*. Let the target interval be denoted as $[start, end]$. We define ToTG as:

$$[start, end] = \Gamma(V, Q_{\text{task}}),$$

where Γ is a grounding function that selects the video segment containing the necessary task-related visual cues, even though such cues are not directly mentioned in the query.

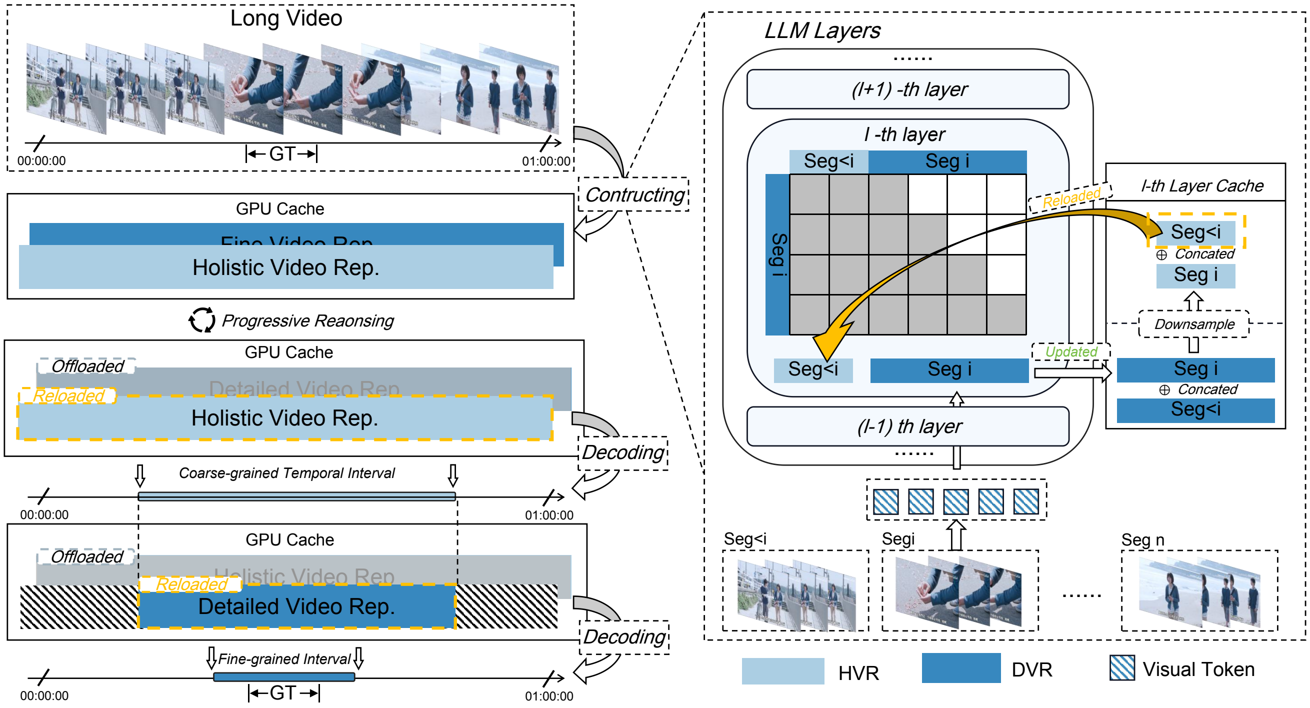


Figure 2. Overview of TimeScope. The input long video is processed to generate two representations: the Holistic Video Representation (HVR), which captures global context, and the Fine Video Representation (FVR), which retains detailed local information. TimeScope first performs coarse-grained reasoning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

This formulation represents a broad variety of real-world scenarios (such as instructional videos, plot reasoning, and surgical workflows) where the required visual evidence is implicit and must be located in the video rather than retrieved through explicit semantic matching.

4. Method

This section outlines the benchmark **ToTG-Bench**, the proposed framework **TimeScope**, and the specialized dataset **ToTG-Pile** in Sec. 4.1, Sec. 4.2, and Sec. 4.3, respectively.

4.1. ToTG-Bench

To enable a comprehensive evaluation of ToTG, we introduce **ToTG-Bench**. We construct ToTG-Bench based on long-video QA datasets, since their most questions naturally align with the definition of ToTG—queries that implicitly specify task-required information within extended video contexts. Moreover, such datasets cover a wide range of domains and task types, allowing us to build a benchmark that is both diverse and realistic. Concretely, we collect question–answer pairs from four public long-video understanding datasets [5, 10, 39, 47], using the questions as task instruction style queries Q_{task} and deriving the corresponding ground-truth temporal intervals through a semi-automated annotation pipeline. First, Gemini-2.5-Pro is used to predict multiple temporal intervals as candidate segments. These candidates are then carefully verified, selected, and refined manually to produce the final grounded annotations. As shown in Table 1, ToTG-Bench exhibits substantially higher diversity and comprehensiveness compared with both traditional temporal grounding benchmarks (Charades-STA [33], ActivityNet [2], and V-STaR [5]) and clue-grounded QA benchmarks (CG-Bench [3] and Next-GQA [40]). It covers a wide range of video durations—from seconds to over one hour—with an average length of 13.5 minutes, and spans 35 realistic video domains (e.g. vlogs, news, documentaries, and sports). Moreover, ToTG-Bench incorporates Q_{task} from both perception tasks (e.g., action or object recognition) and reasoning tasks (e.g., temporal or causal reasoning), covering a total of 12 distinct task types. These characteristics make ToTG-Bench a diverse, comprehensive, and realistic foundation for benchmarking task-oriented temporal grounding models. More details about ToTG-Bench can be found Appendix.

4.2. TimeScope

We propose **TimeScope**, a progressive reasoning framework that reduces the search space in a coarse-to-fine manner: it first identifies a coarse temporal interval where the target is likely to occur, and then performs refined grounding within this narrowed region.

Holistic & Detailed Representation. Dense frame sampling is crucial for accurate grounding in long videos, as it preserves fine-grained temporal continuity and reduces the risk of missing brief task-critical moments. However, existing methods such as UniTime [22] are limited to sparse sampling due to the prohibitive memory cost of processing long sequences. To reconcile dense sampling with memory constraints, TimeScope decouples global context from local details. Specifically, we define a *Holistic Video Representation (HVR)* to abstract long-range context with minimal memory overhead, and a *Detailed Video Representation (DVR)* to retain high-resolution visual cues for precise localization.

To avoid the memory overflow of processing all frames simultaneously, TimeScope constructs both representations in a streaming manner. As illustrated in Fig. 2 (right), the densely sampled video is divided into temporal segments $S_0, S_1, \ldots, S_n$. Each segment S_i is sequentially processed by the MLLM to produce both representations. For each layer l, the key–value states computed from S_i form its detailed representation DVR_i^l. These KVs are then temporally downsampled to obtain the holistic counterpart HVR_i^l, which summarizes segment-level semantics at significantly reduced resolution.

After that, all previously generated holistic caches $HVR_j^l \mid j < i$ are concatenated to form a lightweight historical memory. The current detailed representation DVR_i^l attends to this holistic memory through cross-attention, enabling efficient integration of long-range temporal information without incurring full computation on the entire video. The resulting updated states are forwarded to obtain DVR_i^{l+1}. After all segments are processed, the per-segment representations DVR_i and HVR_i are concatenated along the temporal dimension to form the final detailed- and holistic-level summaries of the entire video, which are stored for subsequent progressive reasoning.

Progressive Reasoning. With both HVR and FVR prepared, TimeScope performs grounding in a coarse-to-fine manner. In the first step, only the compact holistic representation HVR is kept in GPU memory, while the high-resolution FVR remains stored in CPU memory. Leveraging the lightweight long-range context encoded in HVR, the MLLM efficiently predicts a coarse temporal interval that is most likely to contain the task-relevant moment for the query Q_{task}. This stage effectively eliminates the majority of irrelevant frames and substantially reduces the temporal search space. Next, TimeScope reloads only the fine-grained FVR corresponding to the predicted interval and performs detailed reasoning. The rich local temporal and visual details preserved in FVR allow the model to refine the boundaries within the coarse region and accurately localize the target moment. By combining the global efficiency of HVR with the local precision of FVR, the progressive reasoning process achieves high localization accuracy while maintaining low computational cost of processing the long video at high resolution.

4.3. ToTG-Pile

To maximize TimeScope's capacity for task-oriented temporal grounding, we created **ToTG-Pile**—a large-scale dataset bridging conventional temporal grounding data with Task-oriented Temporal Grounding requirements. ToTG-Pile features two key characteristics: (1) all samples follow task-oriented query definitions (Sec. 3), where queries are formulated as task instructions rather than explicit temporal descriptions, and (2) each sample includes both ground-truth temporal intervals and chain-of-thought (CoT) annotations that detail the localization reasoning process.

We constructed the dataset by collecting diverse videos with broad coverage of tasks and visual contexts. Specifically, we leverage comprehensive Video VQA datasets such as VideoR1 [9] and FineVideo [8]. To curate high-quality reasoning data, we employ a three-stage pipeline: (1) **Answer-aware prompting**: we feed answers from original VQA samples into expert temporal grounding models to obtain candidate temporal intervals; (2) **Cross-validation filtering**: we leverage multiple expert models for cross-validation, discarding low-quality samples with IoU < 0.1 across models; (3) **CoT annotation generation**: we use reasoning-capable MLLMs to generate chain-of-thought annotations that detail the localization reasoning process. Additionally, to extend the dataset to long-video scenarios, we synthesized 90K long clips (10 minutes each) by concatenating short videos and generated corresponding temporal grounding queries from video captions. This augmentation enhances TimeScope's capability in handling both conventional temporal grounding and task-oriented temporal grounding in extended video contexts.

Overall, ToTG-Pile provides a large-scale, diverse, and reasoning-oriented foundation for training MLLMs to perform effective and generalizable task-oriented temporal grounding in long videos.

5. Experiment

5.1. Implementation details

We adopt VideoXL-2 [29] as our backbone for two reasons: (1) it can process very long video sequences, enabling straightforward construction of long-video temporal understanding methods, and (2) its internal design interleaves timestamp tokens, providing the model with a strong built-in temporal awareness. **Stage 1: Basic Localization.** We use the temporal grounding splits of ToTG-Pile and train the model to predict target time intervals directly from raw

Method	S(R1@0.3)	S(R1@0.5)	M(R1@0.3)	M(R1@0.5)	L(R1@0.3)	L(R1@0.5)
Temporal Grounding Models						
TimeR1-7B [38]	20.1	16.0	6.8	4.1	15.8	11.0
VideoChatR1-7B [21]	17.3	14.0	6.6	5.6	12.1	10.9
Temporal-RLT-7B [18]	54.0	35.3	17.9	12.3	14.1	11.5
UniTime [23]	52.4	42.7	29.6	23.2	24.2	22.6
Video Understanding Models						
Qwen-2.5VL-7B [1]	48.2	40.6	12.2	10.3	14.6	12.3
Keye-VL-1.5-8B [41]	36.0	34.7	25.4	25.4	32.7	28.6
TimeScope-7B	**52.3**	**46.3**	**45.2**	**42.3**	**47.3**	**37.8**

Table 2. Performance comparison on ToTG-bench. “S” refer to ”Short”, “M” refer to “Medium”, “L” refer to “Long”. “has option” represents incorporating the options into the prompt.

Model	OR	SR	TR	AR	SP	AP	TP	EG	OP	OB	AT	CP	TU
TimeR1-7B	13.8	17.2	11.8	6.50	2.90	14.8	8.80	8.20	7.70	17.4	7.70	13.3	12.5
Temporal-RLT-7B	22.4	20.7	23.5	8.70	**42.9**	48.1	23.5	6.40	26.9	23.9	10.8	40.0	6.20
Qwen-2.5VL-7B	24.1	34.5	23.5	19.6	1.20	40.7	17.6	10.0	30.8	28.3	18.7	33.3	9.30
Keye-VL-1.5-8B	19.0	27.6	35.3	41.3	10.7	22.2	**55.9**	10.0	26.9	30.4	36.9	26.7	**56.2**
TimeScope-7B	**42.9**	**48.3**	**47.1**	**47.8**	36.7	**63.0**	35.3	**30.0**	**46.2**	**42.9**	**36.9**	**40.0**	35.8

Table 3. Performance comparison on different task types in the ToTG-bench. All values are rounded to three decimal places. The metric is IoU@0.5. OR: Object Reasoning, SR: Spatial Reasoning, TR: Temporal Reasoning, AR: Action Reasoning, SP: Spatial Perception, AP: Attribute Perception, TP: Temporal Perception, EG: Ego, OP: OCR Problems, AT: Action Recognition, CP: Counting Problem, TU: Tutorial.

video and task descriptions, bootstrapping its basic localization ability. **Stage 2: Coarse-to-Fine Refinement.** We apply heavy temporal augmentations (random cropping, shifting, and scaling of time spans) to training videos, forcing the model to first estimate a coarse temporal window from abstract video representations and then refine it into fine-grained intervals using detailed representations.

5.2. Results on Task-oriented temporal grounding

Evaluation on ToTG-Bench. We evaluate TimeScope on ToTG-Bench, which categorizes videos into three duration segments: short (<180s), medium (180–600s), and long (>600s). We report IoU@0.3 and IoU@0.5 metrics and compare TimeScope against both specialized temporal understanding MLLMs and general video understanding models. As shown in Table 2, TimeScope demonstrates strong and consistent performance across all video durations. Most notably, TimeScope outperforms all baselines by 20–30 points on medium and long videos, demonstrating substantial advantages in handling task-oriented temporal grounding in extended contexts. On short videos, TimeScope achieves competitive performance, trailing the state-of-the-art Temporal-RLT [18] by less than 2 points while significantly surpassing it on longer videos. These results highlight TimeScope’s effectiveness in task-oriented temporal localization, particularly its ability to perform accurate reasoning and grounding in long-video scenarios while maintaining robust performance across different video durations.

We further evaluate TimeScope and baseline models on ToTG-Bench across different task categories using IoU@0.5 metric, as shown in Table 3. TimeScope demonstrates strong performance across all task categories. Notably, in reasoning tasks (including Object Reasoning, Spatial Reasoning, Temporal Reasoning, and Action Reasoning), TimeScope outperforms all baselines by 10–20 points, highlighting its superior capability in task-oriented reasoning and temporal localization. Across other task categories, TimeScope consistently achieves top-tier performance, demonstrating robust generalization to diverse task types.

5.3. Results on Traditional Temporal Grounding

We conduct a comprehensive comparison of TimeScope against traditional and MLLM-based methods on conventional temporal grounding benchmarks, covering both short-video and long-video settings.

Short-video Benchmarks. As shown in Table 4, TimeScope achieves state-of-the-art performance across all short-video benchmarks. On Charades-STA, TimeScope attains an R1@0.7 score of 64.0, significantly surpassing VideoChat-Flash (27.6), TimeSuite (43.0), and Time-R1 (50.1). On ActivityNet, it achieves an R1@0.7 score of 59.0, outperforming HawkEye (34.7) and Time-R1 (39.0). Notably, TimeScope maintains a smaller gap between R1@0.5 and R1@0.7 compared to most baselines, indicating its capability for more precise temporal localization.

Long-video Benchmarks. As shown in Table 5, TimeScope achieves an R1@0.7 score of 85.2 on V-STaR (with videos up to 300 seconds), substantially exceeding UniTime (62.9) and Keye-1.5-VL (49.1). This demon-

Method	**Charades-STA**			**ActivityNet**		
	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
			Open-source VLP Method			
2D-TAN [46]	45.8	27.9	–	60.4	43.4	–
UniVTG [24]	60.2	38.6	–	56.1	43.4	–
SSRN [48]	65.5	42.6	–	–	54.5	–
SnAG [27]	64.6	46.2	–	–	48.6	–
EaTR [16]	68.4	44.9	–	–	58.2	–
			Open-source MLLMs Method			
TimeChat [32]	32.2	13.4	32.2	36.2	20.2	21.8
VTimeLLM [13]	27.5	11.4	31.2	44.0	27.8	30.4
VideoChat-Flash [19]	53.1	27.6	–	–	–	–
TRACE [12]	61.7	41.4	41.4	37.7	24.0	39.0
HawkEye [37]	58.3	28.8	–	55.9	34.7	–
TimeSuite [44]	67.1	43.0	–	–	–	–
Time-R1 [38]	72.2	50.1	–	58.6	39.0	–
DeepVideo-R1-7B [28]	71.7	50.6	**61.2**	33.9	18.0	36.9
VideoChat-R1-7B [21]	71.7	50.2	60.8	33.4	17.7	36.6
TimeZero-7B [38]	60.8	35.3	58.1	39.0	21.4	40.5
Temporal-RLT-7B[18]	67.9	44.1	57.0	38.4	20.2	39.0
TimeScope-7B	**78.9**	**61.2**	56.2	**66.9**	**56.0**	**46.0**

Table 4. Performance comparison on short video temporal grounding tasks including Charades-STA and ActivityNet.

Model	**R1@0.5**	**R1@0.7**
Qwen2.5-VL-7B	0.0	0.0
UniTime	62.9	62.9
Keye-1.5-VL-8B	63.9	49.1
TimeScope-7B	**87.5**	**85.2**

Table 5. Performance comparison on long video temporal grounding benchmark V-StaR (duration>300).

strates TimeScope's strong capability in handling extended temporal grounding tasks.

Method	**CG-Bench** **Acc.**	**MLVU** **Acc.**	**LongVideoBench** **Acc.**
Uniform Sample	33.87	60.53	54.82
UniVTG [24]	34.87	62.56	54.67
VTimeLLM [13]	34.60	59.52	54.30
TimeSuite [44]	32.47	58.51	53.25
UniTime-Full [23]	**40.30**	66.50	56.47
TimeScope-7B	38.47	**68.12**	**58.34**

Table 6. Performance comparison on Long Video Understanding tasks including CG-Bench, MLVU and LongVideoBench.

5.4. Benefits to Long-Video Understanding

As discussed in Section 3, TimeScope's strong performance on task-oriented temporal grounding demonstrates its potential to help MLLMs capture critical information in long videos for question answering. To validate this, we conduct experiments where TimeScope and baseline temporal grounding models first localize relevant time intervals, and then feed frames from the predicted intervals into Qwen2-VL-7B [1] for answer generation. We compare these results against other grounding models and a default uniform sampling baseline without temporal grounding.

We evaluate on three long-video understanding benchmarks: CG-Bench, MLVU, and LongVideoBench. As shown in Table 6, TimeScope demonstrates strong performance and brings substantial improvements over uniform sampling across all benchmarks, surpassing most temporal grounding baselines. On MLVU and LongVideoBench, TimeScope achieves the highest scores of 68.12 and 58.34 respectively, demonstrating its effectiveness in identifying task-relevant temporal segments for long-video understanding. Notably, while many video understanding questions require information beyond a single temporal segment, TimeScope's task-oriented grounding still provides meaningful performance gains, validating its practical utility for downstream applications.

5.5. Ablation Studies

Effectiveness of Progressive Reasoning. To evaluate the effectiveness and necessity of progressive reasoning in long-video scenarios, we conducted an ablation study comparing two settings: progressive reasoning versus standard single prediction. The experiment was performed on samples longer than 300 seconds from the V-STaR benchmark and samples exceeding 600 seconds from the ToTG-Bench, with results presented in Table 7. The findings demonstrate that progressive reasoning achieves substantial improvements over the single-step baseline across both traditional temporal grounding tasks and task-oriented tasks in long videos. Particularly notable is the performance gain

Method	V-STaR(long)		ToTG-bench(long)	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7
Zeroshot	73.2	61.4	37.8	29.5
Progressive Reasoning	**86.4**	**83.0**	**37.8**	**34.0**

Table 7. Comparison of progressive reasoning versus standard prediction on long-video subsets of V-STaR (duration > 300s) and ToTG-Bench.

Method	Frames	Prefill	Decode	Sum
w/o TimeScope	800	2903ms	816ms	3719ms
w TimeScope	800 1200	2259ms 3205ms	517ms 627ms	2776ms 3832ms

Table 8. The efficiency of TimeScope.

huge improvement in IoU@0.7, which confirms the effectiveness of progressively narrowing the search space for achieving precise temporal localization.

Efficiency of Progressive Reasoning. Despite employing progressive reasoning to achieve higher prediction accuracy, TimeScope remains highly efficient—thanks to its streaming-based video representation construction and carefully designed dual-level representation. As shown in Table 8, our framework enables the MLLM backbone to process significantly more frames while simultaneously improving throughput in both the prefill and decode stages, outperforming baseline methods without TimeScope in both speed and scalability.

Effectiveness of Holistic Video Representation. As outlined in Sec. 4.2, TimeScope enables dense frame sampling for long videos—made feasible by the efficient design of its Holistic Video Representation (HVR). As demonstrated in Table 9a, with HVR, TimeScope supports dense sampling even for extremely long videos (up to 2,000 frames). This capability directly translates into significant gains in temporal grounding precision, particularly for fine-grained event localization.

Efficiency of ToTG-pile Dataset. To further analyze the impact of the ToTG-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 ToTG-bench (long), as shown in Table 9b. It can be observed that the performance gap becomes particularly pronounced on V-STaR and ToTG-bench (long), where ToTG-pile brings performance gains of 40 and 20 points, respectively, compared to the setting without ToTG-pile. This demonstrates the crucial value of ToTG-pile in establishing TimeScope's long-video understanding capability and task-oriented reasoning proficiency.

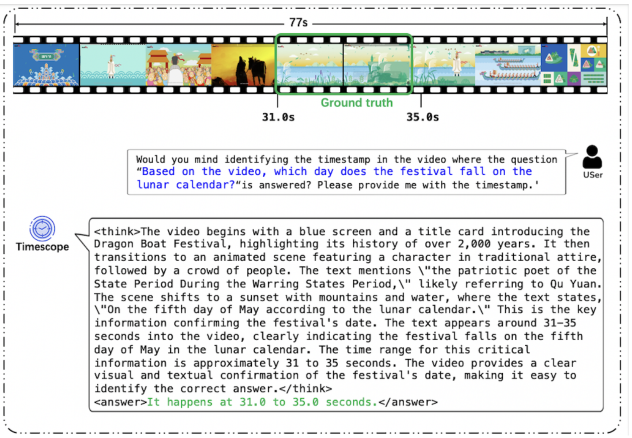
Setting	Max Frames	V-STaR(long)		ToTG-Bench(long)	
		R1@0.5	R1@0.7	R1@0.5	R1@0.7
w/o Holistic VR.	800	73.2	61.4	37.8	29.5
w Holistic VR.	2000	**80.7**	**76.1**	**41.0**	**37.1**

(a) The effectiveness of Holistic Video Representation.

Method	Charades-STA		V-STaR(long)		ToTG-bench(long)	
	R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7
W/o ToTG-Pile	74.9	55.8	47.7	39.7	29.4	22.7
W ToTG-Pile	**78.9**	**61.2**	**86.4**	**83.0**	**37.8**	**34.0**

(b) Analysis of training effect from ToTG-Pile.

Table 9. Ablation studies on model components.



Would you mind identifying the timestamp in the video where the question "Based on the video, which day does the festival fall on the lunar calendar?"is answered? Please provide me with the timestamp.'

<think>The video begins with a blue screen and a title card introducing the Dragon Boat Festival, highlighting its history of over 2,000 years. It then transitions to an animated scene featuring a character in traditional attire, followed by a crowd of people. The text mentions \"the patriotic poet of the State Period During the Warring States Period,\" likely referring to Qu Yuan. The scene shifts to a sunset with mountains and water, where the text states, \"On the fifth day of May according to the lunar calendar.\" This is the key information confirming the festival's date. The text appears around 31-35 seconds into the video, clearly indicating the festival falls on the fifth day of May in the lunar calendar. The time range for this critical information is approximately 31 to 35 seconds. The video provides a clear visual and textual confirmation of the festival's date, making it easy to identify the correct answer.</think>
<answer>It happens at 31.0 to 35.0 seconds.</answer>

Figure 3. Visualization of TimeScope.

6. Qualitative results

We show the results of TimeScope in Figure 3, which can be seen that TimeScope exhibits robust performance with good coarse-grained segment retrieval and fine-grained temporal grounding capabilities. More results can be seen in Appendix.

7. Conclusion

In this work, we define a new task—Task-Oriented Temporal Grounding (ToTG)—and formally conceptualize the aforementioned challenges. To foster research in this emerging area, we introduce ToTG-Bench, a benchmark designed to evaluate temporal grounding performance on diverse, real-world, long-form video-understanding scenarios. To tackle these challenges, we propose TimeScope, a novel framework that solves ToTG through step-by-step reasoning. To strengthen TimeScope, we release ToTG-Pile, a dataset expressly engineered to optimize MLLMs for task-oriented temporal grounding. Harvested from diverse real-world long-video corpora and annotated via a carefully engineered pipeline, ToTG-Pile provides large-scale, high-quality training data. Extensive experiments across a wide spectrum of settings show that TimeScope achieves substantial improvements over existing methods on both traditional benchmarks and ToTG-Bench. We hope this work will stimulate future research on Task-Oriented Temporal Grounding and propel MLLMs toward deeper temporal understanding of video.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 2, 3, 6, 7

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 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, 2015. 2, 4

[3] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024. 4

[4] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, Tyler Poon, Max Ehrlich, Tuomas Rintamaki, Tyler Poon, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models, 2025. 3

[5] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025. 4

[6] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning, 2025. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 3

[8] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo, 2024. 5

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 5

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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, 2025. 4

[11] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. *arXiv preprint arXiv:2410.01615*, 2024. 3

[12] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 3, 7

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 7

[14] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 3

[15] De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025. 3

[16] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856, 2023. 7

[17] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 3

[18] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency, 2025. 6, 7

[19] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 7

[20] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling, 2025. 3

[21] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning, 2025. 6, 7

[22] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025. 3, 5

[23] Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal

grounding with generative multi-modal large language models, 2025. 6, 7

[24] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 7

[25] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 3

[26] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023.

[27] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18930–18940, 2024. 3, 7

[28] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J. Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo, 2025. 7

[29] Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification, 2025. 3, 5

[30] Minghao Qin, Yan Shu, Peitian Zhang, Kun Lun, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Task-aware kv compression for cost-effective long video understanding. *arXiv preprint arXiv:2506.21184*, 2025. 3

[31] 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. Learning transferable visual models from natural language supervision, 2021. 3

[32] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 3, 7

[33] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 4

[34] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. 3

[35] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering, 2024. 3

[36] Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Li, Jose M Alvarez, Lei Zhang, and Zhiding Yu. Videoitg: Multimodal video understanding with instructed temporal grounding. *arXiv preprint arXiv:2507.13353*, 2025. 3

[37] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024. 7

[38] 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. Time-r1: Post-training large vision language model for temporal video grounding, 2025. 3, 6, 7

[39] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 2, 4

[40] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 4

[41] 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, Jiaming Huang, Jiangxia Cao, Jiankang Chen, Jingyun Hua, Kaibing Chen, Kaiyu Jiang, Kaiyu Tang, Kun Gai, Muhao Wei, Qiang Wang, Ruitao Wang, Sen Na, Shengnan Zhang, Siyang Mao, Sui Huang, Tianke Zhang, Tingting Gao, Wei Chen, Wei Yuan, Xiangyu Wu, Xiao Hu, Xingyu Lu, Yi-Fan Zhang, Yiping Yang, Yulong Chen, Zeyi Lu, Zhenhua Wu, Zhixin Ling, Zhuoran Yang, Ziming Li, Di Xu, Haixuan Gao, Hang Li, Jing Wang, Lejian Ren, Qigen Hu, Qianqian Wang, Shiyao Wang, Xinchen Luo, Yan Li, Yuhang Hu, and Zixing Zhang. Kwai keye-vl 1.5 technical report, 2025. 2, 6

[42] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 3

[43] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024. 3

[44] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving MLLMs for long video understanding via grounded tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 7

[45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 3

[46] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 7

[47] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025. 2, 4

[48] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, and Zeyu Xiong. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. 7

[49] 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, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 3

Overview of Appendix

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. Holistic Video Representation Detail

We conducted a comprehensive ablation on the parameters of Holistic Video Representation (HVR), including the chunk size used to partition the video and the compression ratio applied when generating HVR. We obtain HVR from Detail Video Representation through a pooling operation. During training, we randomly select one-third of the samples to be trained with HVR. As shown in Tables 1 and 2, although using HVR leads to a slight performance drop when the video is input at 1 fps with a maximum of 800 frames, we observe clear performance gains when the maximum number of input frames is increased to 2,000. With HVR, the model surpasses the original zeroshot results.

Method		**Max Frames**	**V-STaR(long)**		**ToTG-Bench(long)**	
VideoChunk	HVR Rate		R1@0.5	R1@0.7	R1@0.5	R1@0.7
0	1*	800	73.2	61.4	37.8	29.5
0	2*	800	63.6	63.6	31.2	27.7
0	4*	800	62.5	60.2	30.4	26.4
150	1*	800	61.3	60.2	29.5	26.1
150	2*	800	63.6	63.6	31.8	27.3
150	4*	800	62.5	61.3	30.6	26.9
300	1*	800	60.2	58.0	28.7	24.9
300	2*	800	63.6	63.6	32.4	26.3
300	4*	800	61.3	59.1	28.6	25.8

Table 1. The effectiveness of Holistic Video Representation.’0’ indicates no video partitioning; ’-’ denotes CUDA out of memory (OOM).

Method		**Max Frames**	**V-STaR(long)**		**ToTG-Bench(long)**	
VideoChunk	HVR Rate		R1@0.5	R1@0.7	R1@0.5	R1@0.7
0	1*	800	73.2	61.4	37.8	29.5
0	1*	2000	-	-	-	-
150	1*	2000	62.5	60.2	31.4	26.7
150	2*	2000	80.7	76.1	41.0	37.1
150	4*	2000	**86.4**	**81.8**	43.2	39.7
300	1*	2000	64.7	62.5	33.4	28.2
300	2*	2000	79.5	78.4	40.6	37.3
300	4*	2000	81.8	80.6	42.7	38.8

Table 2. The effectiveness of Holistic Video Representation.’0’ indicates no video partitioning; ’-’ denotes CUDA out of memory (OOM).

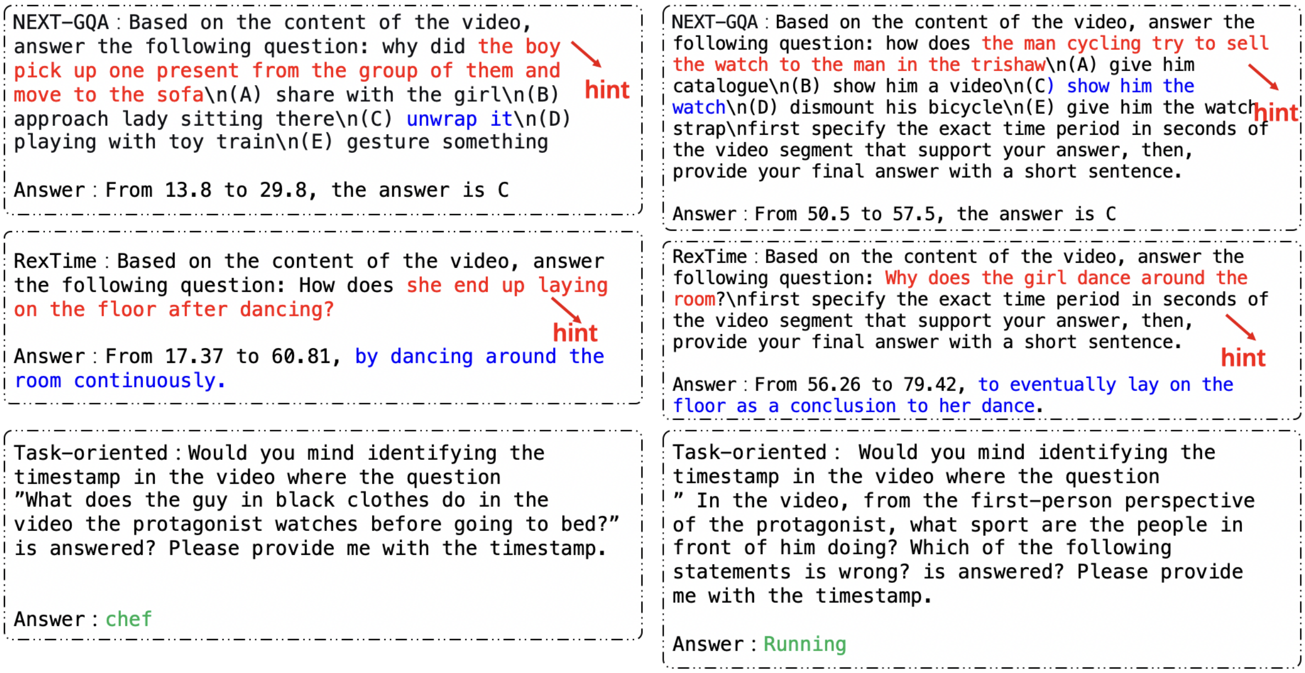
C. ToTG-bench detail

ToTG-Bench comprises 337 videos and nearly 500 questions. As shown in Figure 2, 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 13 task types—such as action reasoning, OCR perception, and temporal reasoning—spanning a wide range of video-understanding tasks.[1] In addition, it includes 35 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.

Explanations of the 13 Task Categories in ToTG-Bench:

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

[1] We adopt the definitions of task type and video category from Video-MME.



NEXT−GQA : Based on the content of the video, answer the following question: why did the boy pick up one present from the group of them and move to the sofa\n(A) share with the girl\n(B) approach lady sitting there\n(C) unwrap it\n(D) playing with toy train\n(E) gesture something

Answer : From 13.8 to 29.8, the answer is C

NEXT−GQA : Based on the content of the video, answer the following question: how does the man cycling try to sell the watch to the man in the trishaw\n(A) give him catalogue\n(B) show him a video\n(C) show him the watch\n(D) dismount his bicycle\n(E) give him the watch strap\nfirst specify the exact time period in seconds of the video segment that support your answer, then, provide your final answer with a short sentence.

Answer : From 50.5 to 57.5, the answer is C

RexTime : Based on the content of the video, answer the following question: How does she end up laying on the floor after dancing?

Answer : From 17.37 to 60.81, by dancing around the room continuously.

RexTime : Based on the content of the video, answer the following question: Why does the girl dance around the room?\nfirst specify the exact time period in seconds of the video segment that support your answer, then, provide your final answer with a short sentence.

Answer : From 56.26 to 79.42, to eventually lay on the floor as a conclusion to her dance.

Task−oriented : Would you mind identifying the timestamp in the video where the question "What does the guy in black clothes do in the video the protagonist watches before going to bed?" is answered? Please provide me with the timestamp.

Answer : chef

Task−oriented : Would you mind identifying the timestamp in the video where the question " In the video, from the first-person perspective of the protagonist, what sport are the people in front of him doing? Which of the following statements is wrong? is answered? Please provide me with the timestamp.

Answer : Running

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.

across frames.

- **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.

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:

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 IoU ≥ 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 IoU ≥ 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 IoU ≥ 0.7

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

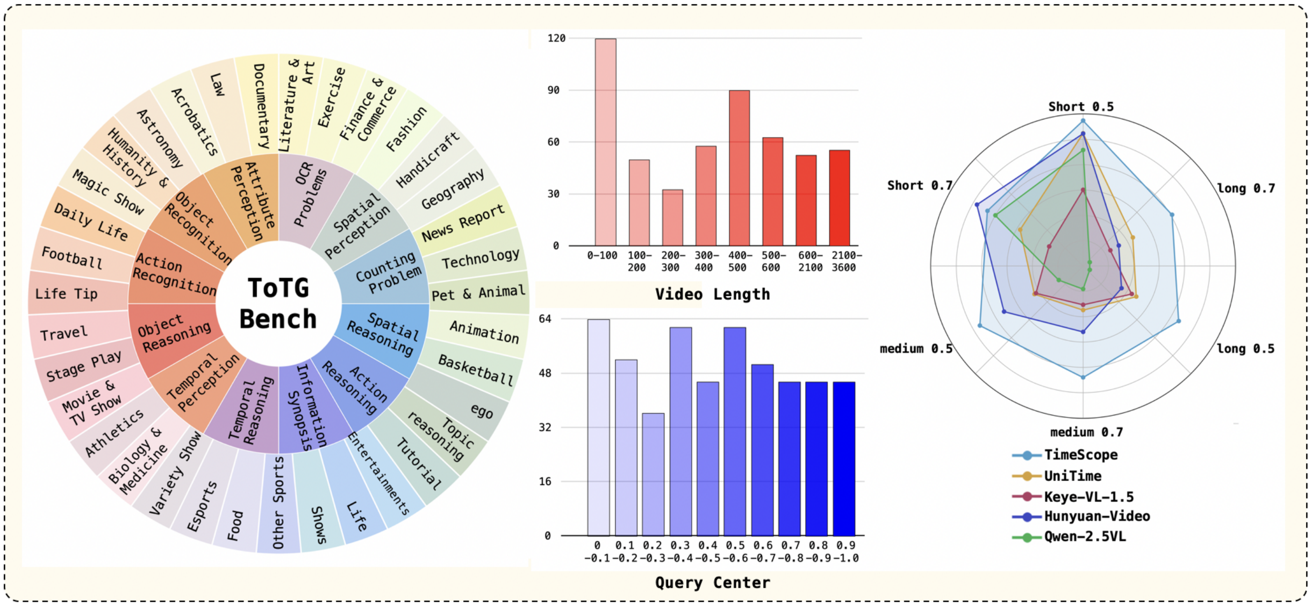


Figure 2. Statistics analysis of ToTG-bench. (Left) Our benchmark covers distinct task types and 35 video categories. (Middle) Video duration and question center distributions. (Right) Performance of various model on ToTG-bench.

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
Qwen2.5VL	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%
Timescope	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%**

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.

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.

D.2. Query Center Robustness

To measure the effectiveness of balanced query centers in the ToTG-benchmark, we evaluated the sensitivity of state-of-the-art (SOTA) models to query 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 Timescope 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 Qwen2.5vl's

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 3. Hyperparameters of Timescope for different training stages

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(<800 frames)

Table 4. Experimental settings of Timescope.

78% and 23.8% higher than UniTime). This further proves the effectiveness of Timescope's training.

E. Experimental Settings & Additional Results

We elaborate on the training and inference details of Timescope. The reported hyperparameters cover Stage 1, and 2, as specified in Table 3. For the inference details, we emphasize the particular context length for different benchmarks, as shown in Table 4.

F. VideoQA Task Details

F.1. VideoQA with Temporal Grounding

We use `Qwen2-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-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

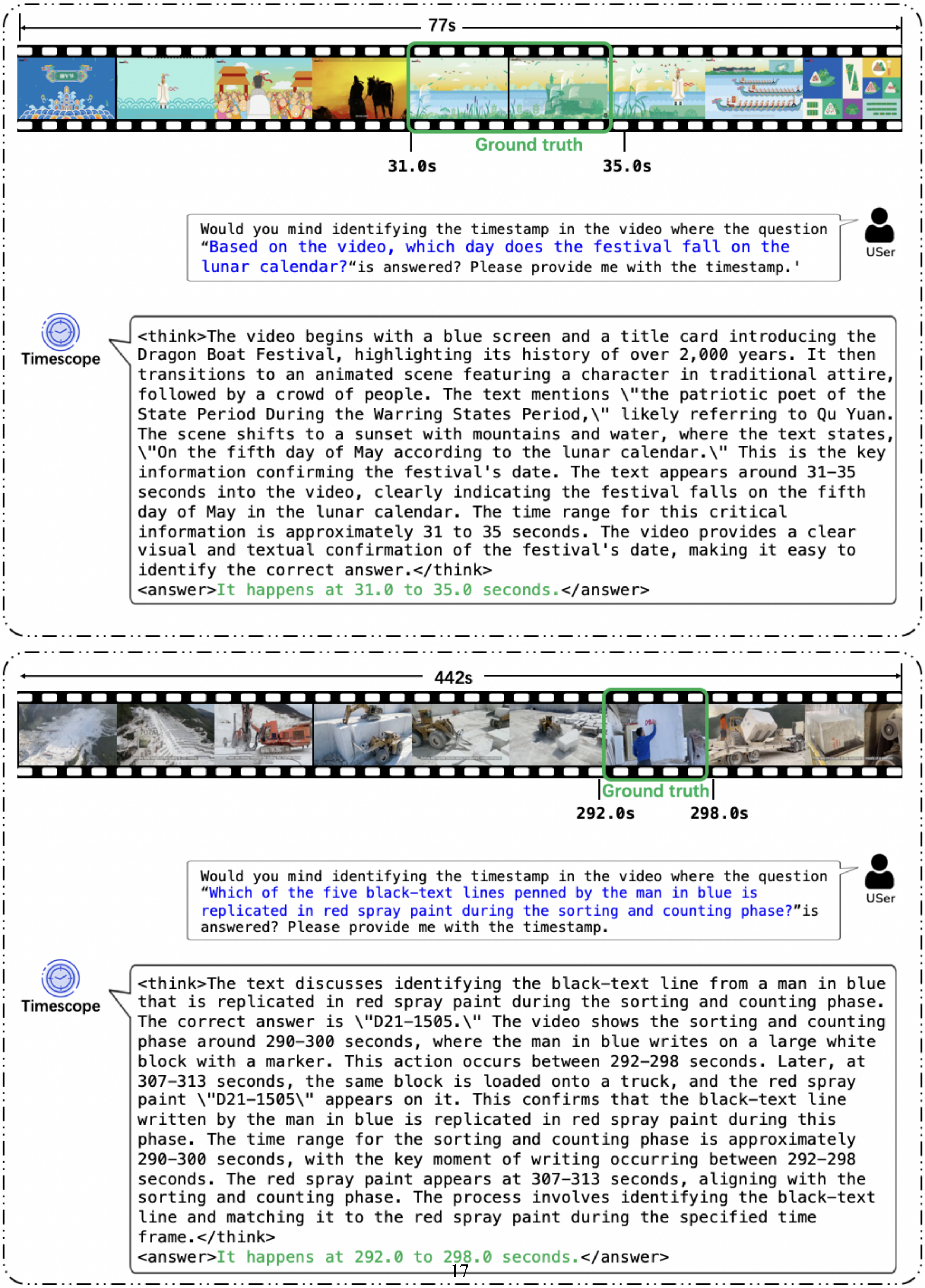
Although Timescope demonstrates exceptional performance on various video temporal grounding and video QA benchmarks, it still has several limitations that warrant further exploration: (i) Timescope is currently constrained to temporal grounding tasks (including traditional temporal grounding tasks and Task-Oriented Temporal Grounding 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 Timescope 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

Our research introduces a new problem, called Task-oriented Temporal Grounding (ToTG), to formally conceptualize the aforementioned challenge. We have created ToTG-Bench, aimed at evaluating temporal grounding performance in diverse real-world long video understanding scenarios and accelerating progress in this emerging field. This facilitates advancing the complexity of temporal understanding in videos and accelerates the development of models that integrate thinking and traditional temporal grounding capabilities. We hope that ultimately, the two can be integrated, unifying problem thinking and temporal grounding. We have also developed a more accurate and efficient temporal grounding framework, Timescope, to advance the field of long video temporal understanding. This could benefit a wide range of downstream applications, such as anomaly detection, security monitoring, etc. We believe that ToTG and Timescope can advance the development of video temporal understanding.

I. Qualitative Results

In Figure 5–7, we present the qualitative results of Timescope on ToTG-bench and V-STaR. Timescope 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 Timescope exhibits robust performance with good coarse-grained segment retrieval and fine-grained temporal grounding capabilities.



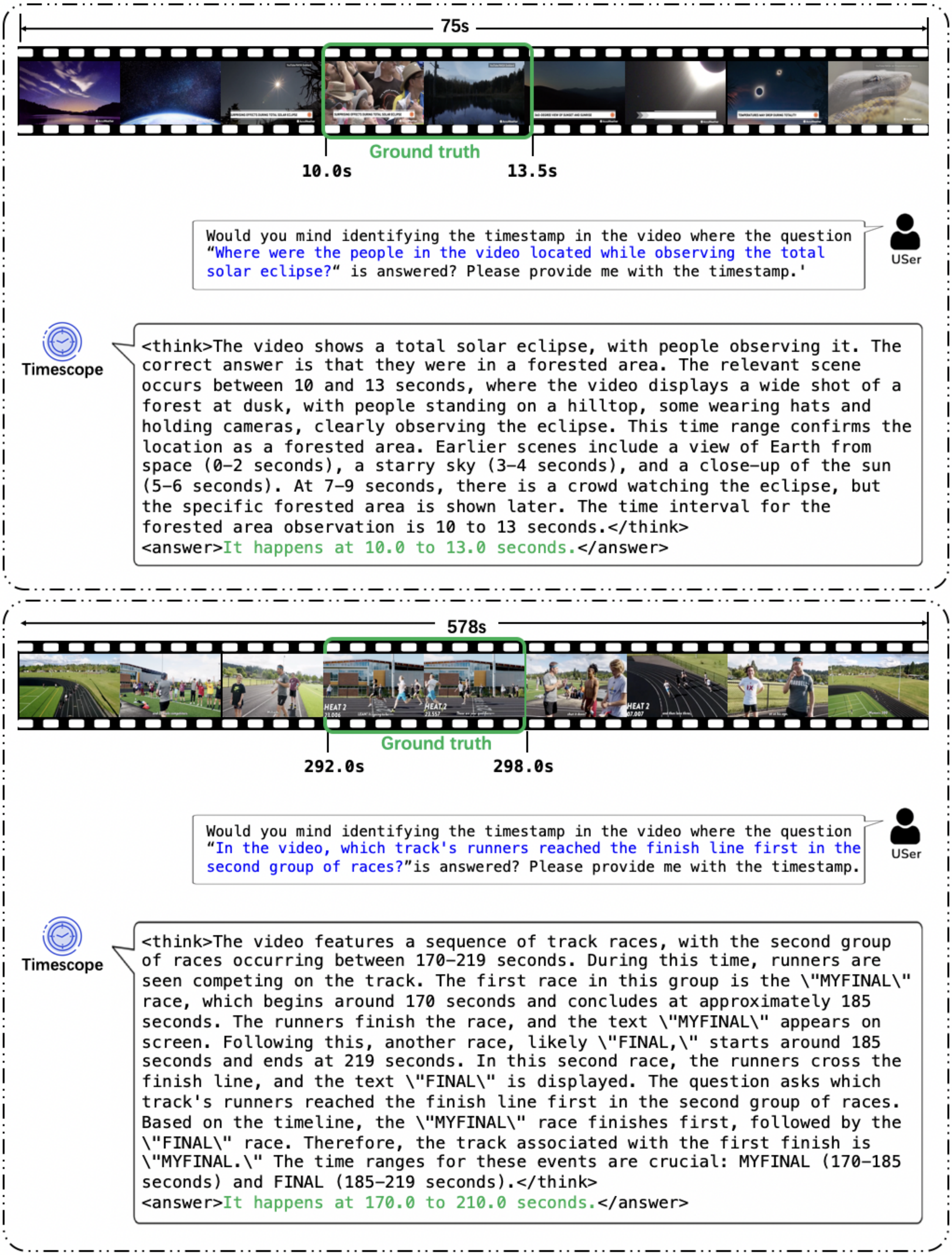
Would you mind identifying the timestamp in the video where the question "Based on the video, which day does the festival fall on the lunar calendar?"is answered? Please provide me with the timestamp.'

<think>The video begins with a blue screen and a title card introducing the Dragon Boat Festival, highlighting its history of over 2,000 years. It then transitions to an animated scene featuring a character in traditional attire, followed by a crowd of people. The text mentions \"the patriotic poet of the State Period During the Warring States Period,\" likely referring to Qu Yuan. The scene shifts to a sunset with mountains and water, where the text states, \"On the fifth day of May according to the lunar calendar.\" This is the key information confirming the festival's date. The text appears around 31-35 seconds into the video, clearly indicating the festival falls on the fifth day of May in the lunar calendar. The time range for this critical information is approximately 31 to 35 seconds. The video provides a clear visual and textual confirmation of the festival's date, making it easy to identify the correct answer.</think>
<answer>It happens at 31.0 to 35.0 seconds.</answer>

Would you mind identifying the timestamp in the video where the question "Which of the five black-text lines penned by the man in blue is replicated in red spray paint during the sorting and counting phase?"is answered? Please provide me with the timestamp.

<think>The text discusses identifying the black-text line from a man in blue that is replicated in red spray paint during the sorting and counting phase. The correct answer is \"D21-1505.\" The video shows the sorting and counting phase around 290-300 seconds, where the man in blue writes on a large white block with a marker. This action occurs between 292-298 seconds. Later, at 307-313 seconds, the same block is loaded onto a truck, and the red spray paint \"D21-1505\" appears on it. This confirms that the black-text line written by the man in blue is replicated in red spray paint during this phase. The time range for the sorting and counting phase is approximately 290-300 seconds, with the key moment of writing occurring between 292-298 seconds. The red spray paint appears at 307-313 seconds, aligning with the sorting and counting phase. The process involves identifying the black-text line and matching it to the red spray paint during the specified time frame.</think>
<answer>It happens at 292.0 to 298.0 seconds.</answer>

Figure 3. Qualitative Results of Timescope.



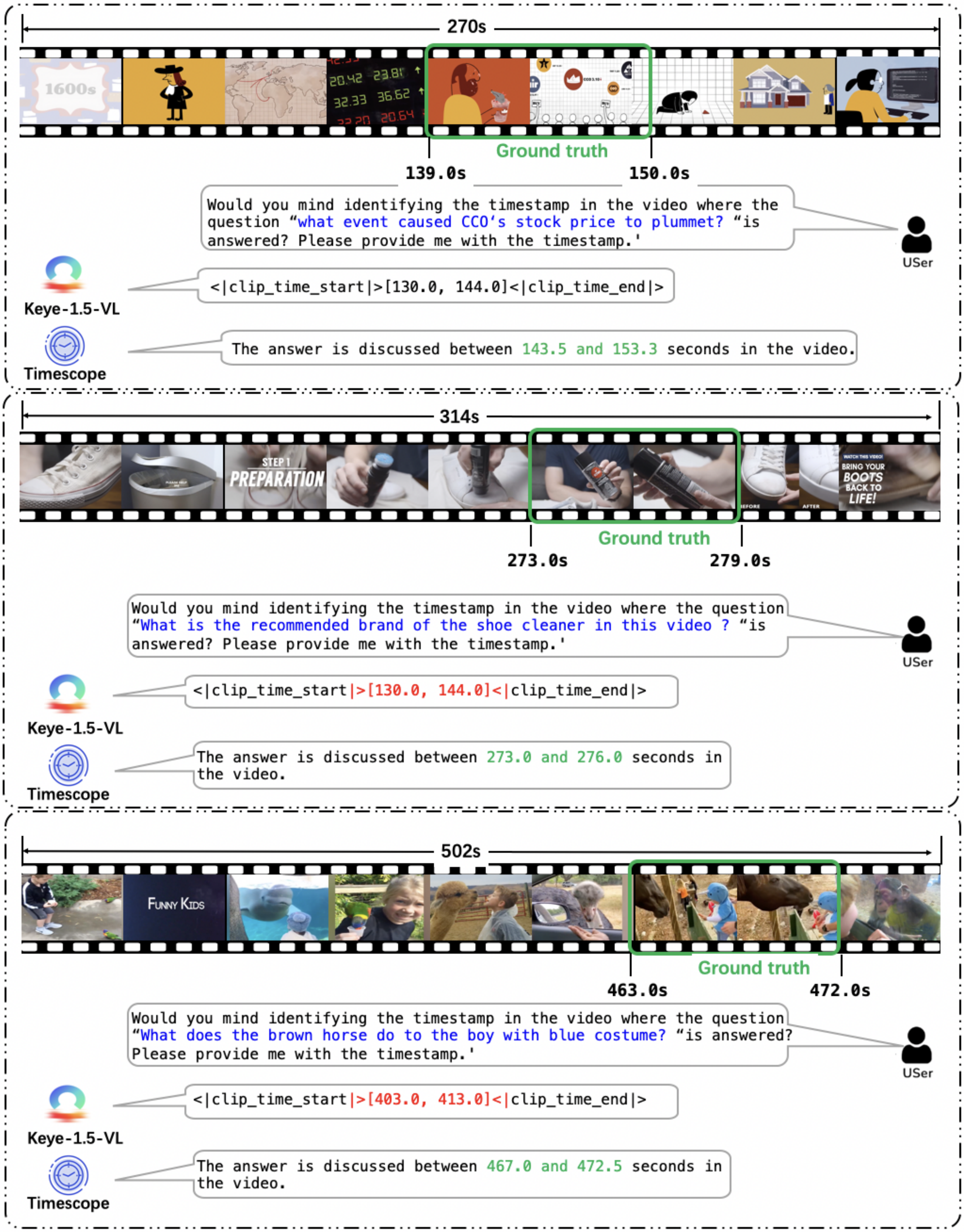
Would you mind identifying the timestamp in the video where the question “Where were the people in the video located while observing the total solar eclipse?“ is answered? Please provide me with the timestamp.'

<think>The video shows a total solar eclipse, with people observing it. The correct answer is that they were in a forested area. The relevant scene occurs between 10 and 13 seconds, where the video displays a wide shot of a forest at dusk, with people standing on a hilltop, some wearing hats and holding cameras, clearly observing the eclipse. This time range confirms the location as a forested area. Earlier scenes include a view of Earth from space (0–2 seconds), a starry sky (3–4 seconds), and a close-up of the sun (5–6 seconds). At 7–9 seconds, there is a crowd watching the eclipse, but the specific forested area is shown later. The time interval for the forested area observation is 10 to 13 seconds.</think>
<answer>It happens at 10.0 to 13.0 seconds.</answer>

Would you mind identifying the timestamp in the video where the question ”In the video, which track's runners reached the finish line first in the second group of races?”is answered? Please provide me with the timestamp.

<think>The video features a sequence of track races, with the second group of races occurring between 170–219 seconds. During this time, runners are seen competing on the track. The first race in this group is the \"MYFINAL\" race, which begins around 170 seconds and concludes at approximately 185 seconds. The runners finish the race, and the text \"MYFINAL\" appears on screen. Following this, another race, likely \"FINAL,\" starts around 185 seconds and ends at 219 seconds. In this second race, the runners cross the finish line, and the text \"FINAL\" is displayed. The question asks which track's runners reached the finish line first in the second group of races. Based on the timeline, the \"MYFINAL\" race finishes first, followed by the \"FINAL\" race. Therefore, the track associated with the first finish is \"MYFINAL.\" The time ranges for these events are crucial: MYFINAL (170–185 seconds) and FINAL (185–219 seconds).</think>
<answer>It happens at 170.0 to 210.0 seconds.</answer>

Figure 4. Qualitative Results of Timescope.



Would you mind identifying the timestamp in the video where the question “what event caused CCO‘s stock price to plummet? “is answered? Please provide me with the timestamp.'

<|clip_time_start|>[130.0, 144.0]<|clip_time_end|>

The answer is discussed between 143.5 and 153.3 seconds in the video.

Would you mind identifying the timestamp in the video where the question “What is the recommended brand of the shoe cleaner in this video ? “is answered? Please provide me with the timestamp.'

<|clip_time_start|>[130.0, 144.0]<|clip_time_end|>

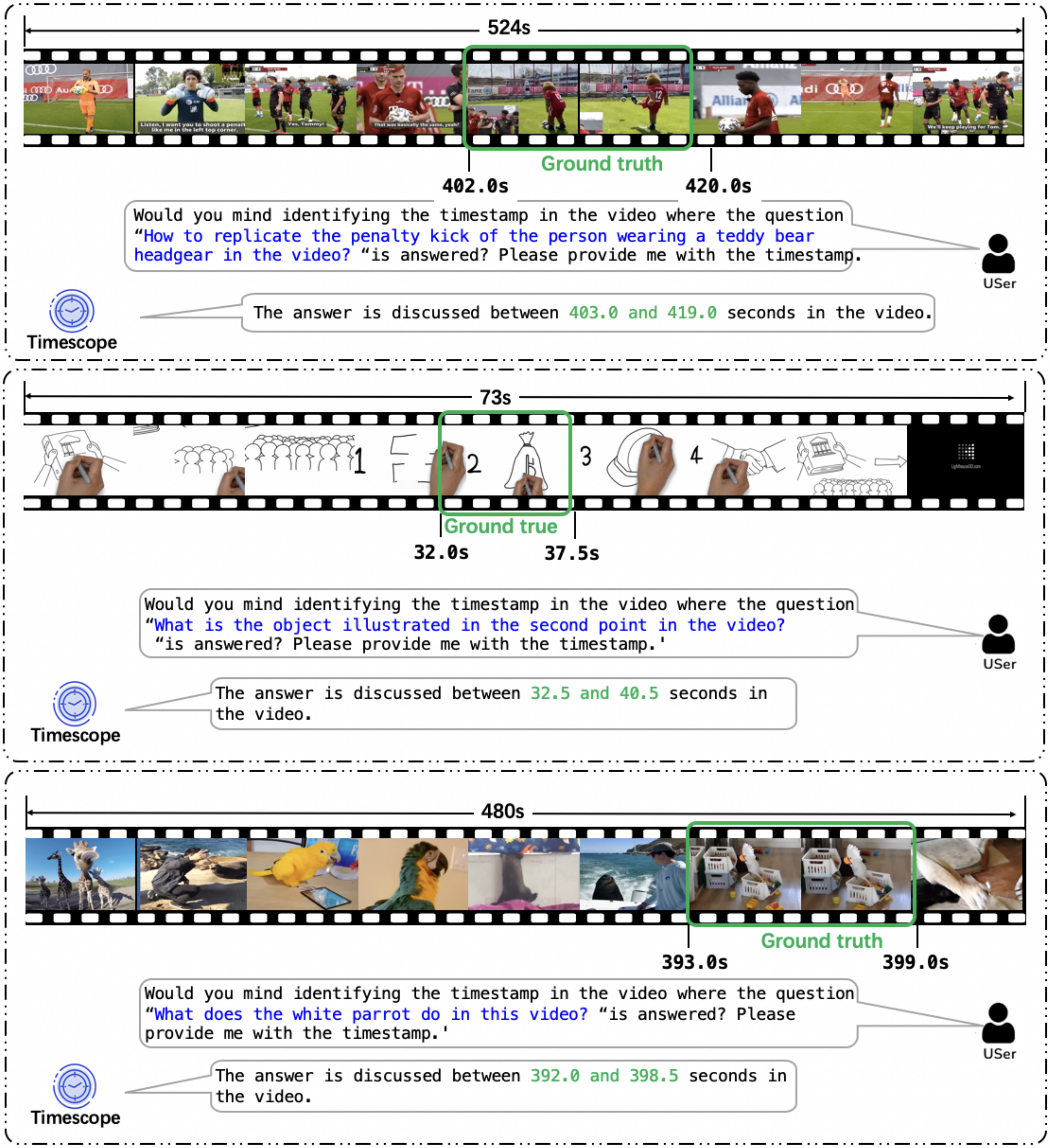
The answer is discussed between 273.0 and 276.0 seconds in the video.

Would you mind identifying the timestamp in the video where the question “What does the brown horse do to the boy with blue costume? “is answered? Please provide me with the timestamp.'

<|clip_time_start|>[403.0, 413.0]<|clip_time_end|>

The answer is discussed between 467.0 and 472.5 seconds in the video.

Figure 5. Qualitative Results of Timescope.



Would you mind identifying the timestamp in the video where the question “How to replicate the penalty kick of the person wearing a teddy bear headgear in the video? “is answered? Please provide me with the timestamp.

The answer is discussed between 403.0 and 419.0 seconds in the video.

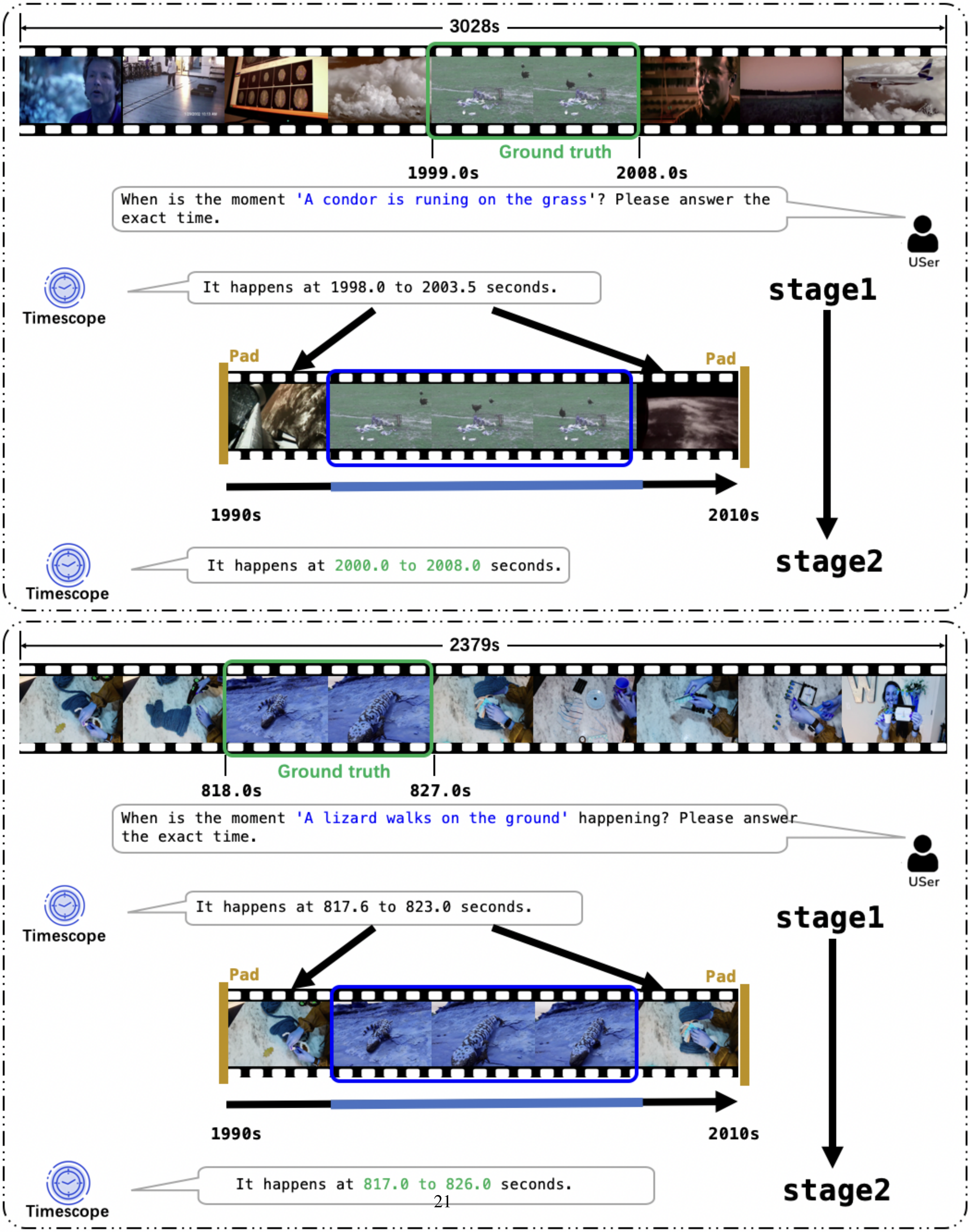
Would you mind identifying the timestamp in the video where the question “What is the object illustrated in the second point in the video? “is answered? Please provide me with the timestamp.'

The answer is discussed between 32.5 and 40.5 seconds in the video.

Would you mind identifying the timestamp in the video where the question “What does the white parrot do in this video? “is answered? Please provide me with the timestamp.'

The answer is discussed between 392.0 and 398.5 seconds in the video.

Figure 6. Qualitative Results of Timescope.



When is the moment 'A condor is runing on the grass'? Please answer the exact time.

It happens at 1998.0 to 2003.5 seconds.

It happens at 2000.0 to 2008.0 seconds.

When is the moment 'A lizard walks on the ground' happening? Please answer the exact time.

It happens at 817.6 to 823.0 seconds.

It happens at 817.0 to 826.0 seconds.

Figure 7. Qualitative Results of Timescope.