

# ZOOMV: TEMPORAL ZOOM-IN FOR EFFICIENT LONG VIDEO UNDERSTANDING

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
Anonymous authors  
Paper under double-blind review

## ABSTRACT

Long video understanding poses a fundamental challenge for large video-language models (LVLMs) due to the overwhelming number of frames and the risk of losing essential context through naive downsampling. Inspired by the way humans watch videos on mobile phones, *constantly zooming in on frames of interest*, we propose **ZoomV**, a query-aware temporal zoom-in framework designed for efficient and accurate long video understanding. Specifically, ZoomV operates in three stages: (1) Temporal interests grounding: guided by the query, ZoomV retrieves relevant events and their associated temporal windows as candidates. (2) Event interests spotlighting: within pools of candidate windows, each window is scored through the model itself reflection and filtered accordingly, where higher-confidence windows are more representative. (3) Compact representation: the selected events are encoded and temporally downsampled to preserve critical semantics while significantly reducing redundancy. Extensive experiments demonstrate that ZoomV substantially outperforms prior **video-agent-style** approaches. On temporal grounding, ZoomV unlocks the latent capability of LVLMs, achieving an **11.8% mIoU** gain on Charades-STA. Remarkably, ZoomV further boosts accuracy on LVBench by **9.7%**, underscoring its effectiveness on long-video benchmarks.

## 1 INTRODUCTION

Long-form video understanding, involving frame sequences that can span minutes to hours, presents a fundamental challenge in computer vision. While large video-language models (LVLMs) have shown impressive performance on video-language tasks (Zhang et al., 2024f; Li et al., 2024b; Caba Heilbron et al., 2015; Yu et al., 2019; Lei et al., 2021; Wu et al., 2024), they still struggle with long videos. On one hand, feeding all frames into the model leads to a rapid increase in computational cost and memory usage, making naive full-frame processing infeasible. On the other hand, aggressive temporal downsampling risks discarding critical context, often resulting in severe visual hallucinations. For instance, the advanced LLaVA-Video uniformly samples only 64 frames regardless of video duration, leading to a significant loss of detailed temporal information, especially in hour-long videos (Zhang et al., 2024f). Therefore, it is crucial to identify *a sufficient number of the most relevant frames in a prompt-aware manner* for reliable long-video understanding.

Existing attempts to scale LVLMs to long videos can be mainly grouped into two directions. The first is token sparsification, reducing sequence length by discarding a subset of visual tokens or patches (Li et al., 2023b; Zhang et al., 2024d; Ma et al., 2025; Zhang et al., 2025). While this improves computational efficiency, it inevitably compromises the holistic integrity of frames, and *often leads to noticeable performance degradation*. The second is video-agent approaches, which typically assemble *a pipeline of heterogeneous models*, e.g., (Wang et al., 2024d; 2025b) using EVA-CLIP (Sun et al., 2023) for retrieval, BLIP (Li et al., 2022) for captioning, and GPT-4 (Achiam et al., 2023) for reasoning. Although such modular systems alleviate sequence-length constraints, their reliance on disparate models makes them inefficient, resource-intensive, and non-end-to-end.

To address the limitations of both paradigms, we take inspiration from human cognitive strategies (Sweller, 1994; Zacks & Swallow, 2007; Wu & Xie, 2024; Shen et al., 2024a), particularly the way humans selectively zoom in on relevant visual content, and introduce our method dubbed ZoomV, a query-aware temporal zoom-in framework for efficient long-video understanding. As illustrated in Figure 1 (b), humans review videos broadly to find relevant clues, then gradually focus on more



Figure 1: **Illustration of human-like interaction for long-video understanding.** It divides hour-long videos into manageable sub-events and searches within query-aware segments.

specific sub-events for detailed inspection. Importantly, when the necessary information is not immediately clear, humans may revisit multiple candidate sub-events iteratively, grounding and validating their relevance until they find satisfactory answers. Therefore, ZoomV imitates the behavior and progressively divides the video timeline into coarse-grained events and finer-grained sub-events, enabling efficient human-like search. Specifically, ZoomV unfolds in three progressive stages.

In the first stage, to identify subtle temporal details within promising sub-events accurately, we first need to ground temporal interest. Previous LVLMs (Li et al., 2023a; Zhang et al., 2024f) have incorporated temporal instructions to improve understanding, yet they do not effectively align visual and temporal cues. In contrast, our ZoomV employs a TemporaLink that explicitly embeds temporal information into visual frame representations, enabling LVLMs to precisely associate visual content with corresponding timestamps. Additionally, to alleviate quantization errors introduced by frame sampling, we optimize the absolute timestamp representation, stabilizing temporal learning and enhancing grounding performance. The model retrieves query-relevant events and their associated temporal windows as candidate regions along the video timeline, providing a coarse yet comprehensive coverage of potentially relevant content.

In the second stage, event interests spotlighting, each candidate window is evaluated through the model’s self-reflection mechanism dubbed TemporaLight, which assigns confidence scores to spotlight the most representative windows while filtering out less relevant ones. As humans hierarchically search through time, they continuously reflect on whether a specific sub-event warrants deeper inspection. Similarly, we leverage the self-reflection capability of LVLMs to guide the search process. Recent studies (Lin et al., 2022; Kadavath et al., 2022; Zheng et al., 2023; Zhang et al., 2024b) have demonstrated that LLMs effectively assess their prediction confidence through additional multiple-choice or yes/no questions. Inspired by these findings, we first identify that LVLMs inherently possess a similar self-reflection capability—“*they know what they do not know.*” This selective scoring process ensures that only high-confidence sub-events proceed to the next stage.

In the third stage, compact representation, the spotlighted events are encoded and temporally down-sampled into condensed representations that preserve critical semantics while substantially reducing redundancy. Together, these stages enable ZoomV to progressively zoom in from coarse-grained temporal coverage to fine-grained and compact representations, achieving efficient and accurate long-video understanding. Extensive experiments demonstrate its superior performance across various challenging benchmarks, including VideoMME (Fu et al., 2024), MLVU (Zhou et al., 2024), and LongVideoBench (Wu et al., 2024). Notably, on the highly challenging LVBench dataset with hour-long videos (Wang et al., 2024c), ZoomV achieves new state-of-the-art accuracy, substantially surpassing previous methods. ZoomV also substantially outperforms existing video grounding methods on temporal grounding tasks, such as Charades-STA (Gao et al., 2017), ActivityNet Captions (Caba Heilbron et al., 2015), and ReXTime (Chen et al., 2024a). We conduct comprehensive ablation studies and reveal the sources of improvement. In summary, our contributions are threefold:

- We reveal that LVLMs inherently exhibit strong self-reflection abilities, previously studied mainly in LLMs, which enable reflection-guided prioritization of temporal search.

- 108 • We propose **ZoomV**, a query-aware hierarchical temporal zoom-in framework that mimics  
109 human coarse-to-fine exploration, significantly advancing long-video understanding. For  
110 example, ZoomV improves accuracy on LVBench from 41.8% to 51.5%.
- 111 • We demonstrate that LVLMs possess latent temporal grounding capabilities, which can be  
112 effectively unlocked through our proposed TemporaLink. Despite its simplicity, this design  
113 achieves an 11.8% mIoU gain over state-of-the-art temporal grounding models.

## 115 2 RELATED WORK

### 116 2.1 LONG VIDEO UNDERSTANDING

117 Understanding lengthy videos for LVLMs is challenging due to the need to store and extract information  
118 effectively from numerous frames. One common line involves using language as a bridge to summarize videos into concise captions (Islam et al., 2024; Zhang et al., 2023a), resulting in the omission of vital visual signals. Another widely studied line involves memory-based methods for compressing video features into a limited memory bank, which is achieved by continually updating the memory bank during visual encoding (Song et al., 2024). Memory bank has also been applied to real-time streaming video understanding, potentially enabling an unlimited length of frames while maintaining a constant space footprint (Zhang et al., 2024a). A major drawback of these methods is their oversight of video duration and information density, particularly when utilizing a fixed space for a memory bank. For instance, Flash-VStream compresses both brief 10-second clips and hour-long movies into the same 681 tokens (Zhang et al., 2024a). Besides, these black box methods lack interpretability, as it is hard to verify whether pertinent details are accurately retrieved for reasoning. Another line of work reduces video tokens through visual compression, including spatial token selection (Zhang et al., 2025), temporal frame selection (Yu et al., 2025; Tang et al., 2025), and joint spatio-temporal selection (Shen et al., 2024b). These methods typically serve as pre-processing strategies before LLM inference. In contrast, ZoomV leverages the inherent temporal grounding ability, enabling task-aware and hierarchical selection.

### 136 2.2 TEMPORAL GROUNDING

137 Temporal Grounding (TG) localizes video segments relevant to textual queries Gao et al. (2017);  
138 Hendricks et al. (2018); Lei et al. (2021) or questions Chen et al. (2024a); Xiao et al. (2024). Current  
139 LVLM-based approaches typically rely on heavy customizations to handle continuous time,  
140 such as specialized time tokens Wang et al. (2024a); Guo et al. (2024); Ren et al. (2024), auxiliary  
141 temporal modules Qian et al. (2024); Wang et al. (2024e), or extensive multi-stage reinforcement  
142 training Li et al. (2025a); Wang et al. (2025a). In contrast, ZoomV show that standard LVLMs inherently  
143 possess precise grounding capabilities when properly guided. ZoomV unlocks this potential  
144 through a simple TemporaLink and coarse-to-fine zooming mechanism, achieving state-of-the-art  
145 performance without architectural modifications or sophisticated RL training.

### 147 2.3 VIDEOAGENTS

148 VideoAgent is a LLM agent that understands videos by using customized structured and tools (Wang  
149 et al., 2024d; Hendricks et al., 2018; Wang et al., 2025b). Previous research usually needs one more  
150 model for the agent pipeline. For instance, Wang et al. (2024d) introduces a prompt-driven video QA  
151 agent that employs extra vision-language retrieval models (e.g., CLIP) to ground key frames during  
152 reasoning. Building on this, VideoTree (Wang et al., 2025b) designs a hierarchical tree-style search  
153 by clustering frames with visual features, enabling structured exploration. Both approaches still rely  
154 on caption models to provide frame-level descriptions. In contrast, our ZoomV avoids additional  
155 captioning or grounding models and directly leverages LVLMs to predict continuous temporal  
156 windows. Moreover, the tree-like structure of ZoomV is constructed purely over simple temporal  
157 segments, without requiring extra visual feature models or clustering algorithms. We also highlight two  
158 concurrent works with hierarchical search strategies. UniTime (Li et al., 2025b) exhaustively per-  
159 forms temporal grounding on all video segments, while VideoChat-R1.5 (Li et al., 2025a) employs a  
160 fixed three-step grounding strategy. In contrast, ZoomV designs an *adaptive* search mechanism with  
161 reflection-guided early termination: TemporaLight assesses confidence to avoid exhaustive searches  
and the priority queue enables flexible exploration with backtracking to alternative paths. These

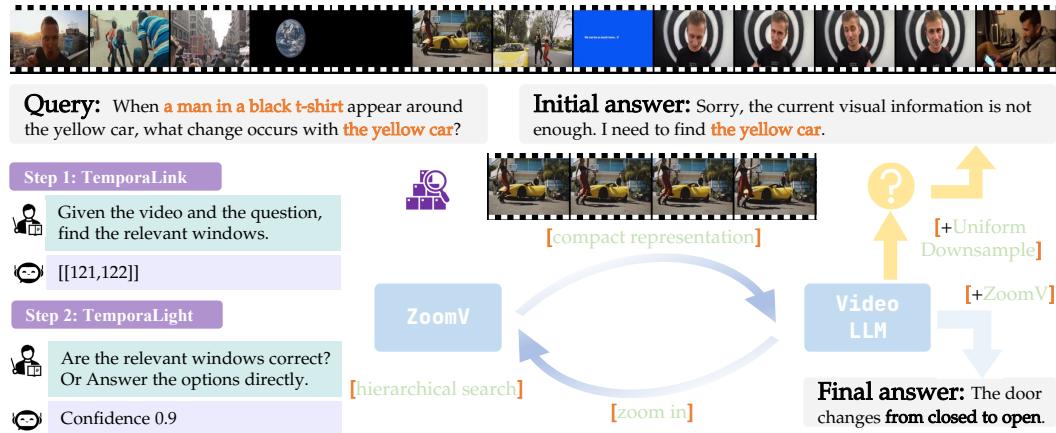


Figure 2: **An illustrative view of ZoomV.** Equipped with ZoomV, Video LLM can gain enhanced capability for efficient and accurate long-video understanding.

innovations allow ZoomV to achieve superior efficiency and accuracy by focusing computation on the most relevant temporal windows.

### 3 PROPOSED APPROACH: ZOOMV

In this section, we present the ZoomV framework, which equips LVLMs with a hierarchical, human-inspired temporal zoom-in mechanism. We first introduce **TemporaLink**, a temporal-augmented representation that explicitly binds timestamps with visual frames to support accurate temporal grounding. Next, we uncover the inherent self-reflection capability of LVLMs and propose **TemporaLight**, and describe the **reflection-guided hierarchical search algorithm**, which integrates grounding and reflection to progressively zoom in from coarse-grained events to fine-grained sub-events. Ultimately, produce compact event representations for efficient long-video understanding.

#### 3.1 PRELIMINARY: UNIFIED AUTOREGRESSIVE MODELING

Our ZoomV is built upon an autoregressive LVLM backbone, which sequentially predicts tokens conditioned on visual and textual contexts. An autoregressive LVLM generates an output sequence  $\mathbf{y} = (y_1, y_2, \dots, y_L)$  with length  $L$  given a text condition  $\mathbf{x}$  and a video condition  $\mathbf{v}$  by predicting tokens one at a time based on the previously generated tokens. Assuming that the LVLM is parameterized by  $\theta$ , the conditional probability distribution of generating a sequence  $\mathbf{y}$  given context  $\mathbf{x}$  and  $\mathbf{v}$  is defined as

$$p_{\theta}(\mathbf{y}|\mathbf{v}, \mathbf{x}) = \prod_{i=1}^L p_{\theta}(y_i|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<i}), \quad (1)$$

where  $\mathbf{y}_{<1} = \emptyset$  and  $\mathbf{y}_{<t} = (y_1, y_2, \dots, y_{t-1})$ . Taking video question answering (VQA) as an example, an LVLM predicts the answer distribution  $p_{\theta}(\mathbf{a} | \mathbf{v}, \mathbf{q}, I_q)$ , where  $\mathbf{q}$  denotes the input question and  $I_q = \text{"Answer the following questions related to this video"}$  serves as the instruction. Here,  $\mathbf{v}$  represents a sequence of  $T$  downsampled frame tokens extracted from the original video, which are transformed by a dedicated visual encoder and projector into visual tokens. In the following sections, we extend this autoregressive formulation to model both the grounding and reflection mechanisms within a unified framework.

#### 3.2 TEMPORAL INTEREST GROUNDING VIA TEMPORALLINK

Temporal interest grounding aims to identify the most relevant temporal windows according to the query, modeling continuous numerical timestamps as discrete digit generations (Ren et al., 2024; Guo et al., 2024). The task process is defined as: (1) Given a query  $\mathbf{q}$  and the grounding instruction  $I_g = \text{"Find the relevant windows"}$ , the model predicts text sequence  $p_{\theta}(\mathbf{w}|\mathbf{v}, \mathbf{q}, I_g)$ ; (2) Then the text sequence  $\mathbf{w}$  is turned into a set of time ranges  $W = [(s_1, e_1), \dots, (s_K, e_K)]$  with

size  $K$ , where  $s_k, e_k$  signifies the start and end timestamps of  $k$ -th target window clip. However, LVLMs naturally struggle to accurately handle numerical tasks, especially in temporal tasks involving precise numerical comparisons (Schwartz et al., 2024; Xie, 2024). To alleviate this challenge and effectively activate the inherent temporal grounding capability of LVLMs, we propose a simple yet effective enhancement module, dubbed **TemporaLink**, which explicitly binds timestamps with visual frame representations.

Specifically, given a downsampled video represented by frames  $(f_1, f_2, \dots, f_T)$  and their corresponding fractional timestamps  $(t_1, t_2, \dots, t_T)$ , e.g.,  $(0.00, 3.33, 6.67, 10.00)$ , we first round these timestamps to the nearest integer. Then, to ensure a consistent token representation in TemporaLink, we apply left-zero padding, resulting in timestamps like  $(00, 03, 07, 10)$ :

$$\tilde{t}_i = \text{Pad}(\text{Round}(t_i)). \quad (2)$$

Next, we extract frame visual features through a visual encoder  $\mathcal{V}$  with a projection module (Zhang et al., 2024f). To embed the absolute timestamp into each frame feature, as illustrated in Figure 3, we directly concatenate these features with their corresponding absolute timestamp embeddings:

$$\tilde{\mathbf{v}}_i = \text{concat}(\mathcal{V}(f_i), \mathcal{T}(\tilde{t}_i)), \quad \tilde{\mathbf{v}}_i \in \mathbb{R}^{(N+P) \times D}, \quad (3)$$

where  $\mathcal{T}$  denotes the embedding layer of the LLM,  $D$  represents the embedding dimension, and  $N, P$  denote the number of visual frame tokens and padded timestamp tokens, respectively. Following the above design, we perform refinement training on the LVLM. During this process, manually annotated timestamps are further aligned with the rounded timestamps to mitigate quantization errors, as detailed in the Appendix B. By explicitly linking visual frames with timestamps, TemporaLink not only significantly enhances the temporal grounding capability of LVLMs but also provides a solid foundation for the subsequent spotlighting and selection stages.

### 3.3 EVENT INTERESTS SPOTLIGHTING VIA TEMPORALIGHT

After obtaining the candidate temporal windows, we need to validate them and highlight the most suitable ones. Previous research has demonstrated that generative LLMs can evaluate the correctness of their predictions through self-reflection mechanisms (Lin et al., 2022; Zheng et al., 2023; Zhang et al., 2024b;e). These models can produce well-calibrated confidence scores for Yes/No and multiple-choice questions. We extend this observation from text-based LLMs to LVLMs and propose TemporaLight to assess the validity of temporal spotlight predictions. Here, there are two forms of validation.

For Yes/No type reflection validation, given a question  $\mathbf{q}$ , a TemporaLight prediction  $W$ , and reflection instruction  $I_{\text{tf}}$  = “Are the proposed relevant windows correct?”, the probability is formulated as

$$c = p_{\theta}(\text{“Yes”} | \mathbf{v}, \mathbf{q}, W, I_{\text{tf}}). \quad (4)$$

The Yes/No reflection confidence positively correlates with grounding accuracy (mIoU), thus providing an intrinsic measure of spotlight correctness without human annotations (Figure 4, top). For multiple-choice reflection validation, the reflection confidence score is defined by selecting the maximum prediction probability from multiple choices. Given a set of candidate answers, the reflection confidence is computed as:

$$c = \max \{p_{\theta}(o | \mathbf{v}, \mathbf{q}, W, I_{\text{mc}})\}, o \in \{\text{“A”, “B”, …}\}, \quad (5)$$

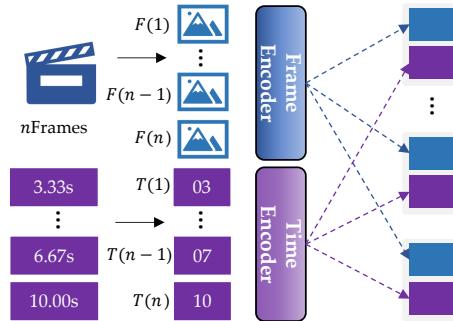


Figure 3: **Illustration of TemporaLink.**

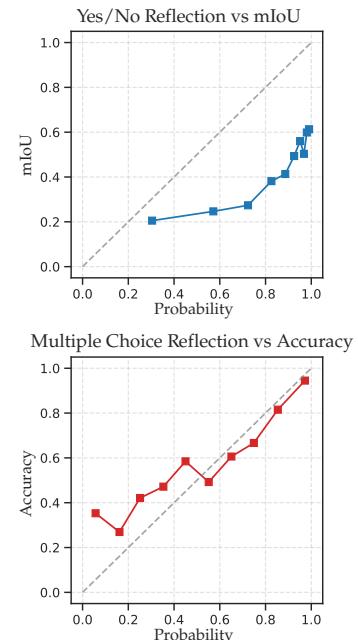


Figure 4: **Reflection probabilities correlate with performance.**



324 

## 4 EXPERIMENTS

325 

### 4.1 EXPERIMENT SETTINGS

326 **Evaluation Benchmarks.** We comprehensively evaluate our ZoomV across three types of video  
 327 understanding tasks, including a total of **eight** subtasks. (1) For **video question answering**, ZoomV  
 328 is validated on three long-video multiple-choice benchmarks, including LongVideoBench (Wu  
 329 et al., 2024), MLVU (Zhou et al., 2024), and LVBench (Wang et al., 2024c), which cover dura-  
 330 tions from minutes to hours, as well as short-video benchmarks like MVBench (Li et al., 2024b)  
 331 and VideoMME (Fu et al., 2024). (2) For **temporal sentence grounding**, we conduct zero-shot  
 332 evaluation on widely used benchmarks such as Charades-STA (Gao et al., 2017) and ActivityNet-  
 333 Captions (Caba Heilbron et al., 2015). Following prior works (Ren et al., 2024; Lin et al., 2023),  
 334 we adopt Recall@1 at IoU thresholds 0.3, 0.5, 0.7 and mIoU as metrics. (3) For **temporal question**  
 335 **grounding** task, we evaluate on the popular ReXTime benchmark (Chen et al., 2024a), which is  
 336 designed to assess temporal reasoning and causal understanding across multiple video events, and  
 337 we measure both VQA accuracy and Recall@1 at IoU thresholds 0.3 and 0.5.

338 **Models.** We apply our ZoomV to the LLaVA-Video (Zhang et al., 2024f), InternVL2.5 (Chen  
 339 et al., 2024b), and advanced Qwen2.5-VL (Team, 2025), three different video model architectures  
 340 for generality. The training of TemporaLink is both completed within eight hours using 128 A100  
 341 GPUs. To enhance TemporaLight capabilities without sacrificing general performance, we apply  
 342 LoRA (Hu et al., 2022) with a rank of 32 to the LLM, and freeze other parameters.

343 **Video Input Format.** Within the identified time window  $W$ , interest frames are densely sampled  
 344 from the spotlighted segments for video understanding. These dense frames are **appended** after the  
 345 globally sparsely sampled frames, thereby retaining the ability to answer questions about the global  
 346 video context. In our experiments, the number of global frames is set to 64, while the maximum  
 347 number of spotlight frames provided by ZoomV is 16.

348 

### 4.2 MAIN RESULTS

349 **Video Question Answering Tasks.** As illustrated in Table 1, ZoomV consistently boosts existing  
 350 open-source LVLMs across a wide range of benchmarks. On short-video tasks such as MLVU,  
 351 ZoomV maintains competitive performance, ensuring that its long-video enhancements do not come  
 352 at the cost of short-duration understanding. The advantage becomes *more pronounced on long-video*  
 353 *datasets*: for instance, ZoomV improves InternVL2.5 on LVBench (average duration 4101 seconds)  
 354 from 41.8% to 51.5%, and **makes it surpass all prior methods**. Similar gains are observed on  
 355 LongVideoBench (+2.7% accuracy) and VideoMME-Long (+1.7%), highlighting its effectiveness  
 356 in handling videos lasting up to several hours. Furthermore, the consistent improvements across  
 357 different LVLM backbones, such as an **8.7%** and **11.3%** boost on LVBench with LLaVA-Video and  
 358 Qwen2.5-VL, demonstrate the robustness and versatility of our ZoomV as a general solution for  
 359 temporal search in long video understanding.

360 **Temporal Grounding Tasks.** As shown in Table 2, ZoomV delivers substantial improvements  
 361 over existing grounding-oriented LVLMs. On standard benchmarks such as Charades-STA and  
 362 ActivityNet Captions, it achieves an average mIoU gain of **11.8%** compared with the previous state-  
 363 of-the-art (e.g., GroundedVideo-LLM). Beyond these datasets, ZoomV demonstrates even stronger  
 364 advantages on ReXTime, where it boosts mIoU by **8.6%**, Recall@0.5 by **9.6%**, and nearly **doubles**  
 365 VQA accuracy from 40.0% to 76.5% relative to TimeChat, clearly highlighting its ability to reason  
 366 over complex temporal event structures.

367 

### 4.3 ANALYSIS

368 **Effectiveness of TemporaLink.** We quantitatively validate TemporaLink by comparing it to the  
 369 widely used Time Instructions on various datasets. As shown in Figure 5a, compared to the time  
 370 instruction, TemporaLink significantly improves the ability of moment retrieval, especially at high  
 371 IoU thresholds. Specifically, when replacing TemporaLink with time instruction on the QVHigh-  
 372 light, R@0.7 drops sharply by 13.8% while R@0.5 drops 6.5%. Although previous LVLMs are  
 373 capable of recognizing relevant events, they struggle to accurately establish associations between

378 **Table 1: Comparison of ZoomV with other LVLMs on video understanding results.** The results  
 379 on various short and long video benchmarks with video durations range from seconds to hours.  
 380

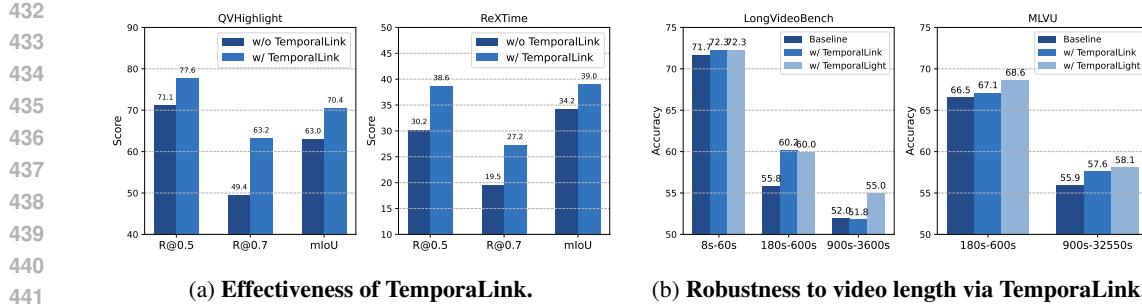
| Model                                   | Size      | #F           | MVBench        | MLVU        | LongVideoBench |             | VideoMME       |             | LVBenchmark    |  |  |  |  |  |
|---|-----------|--------------|----------------|-------------|----------------|-------------|----------------|-------------|----------------|--|--|--|--|--|
|   |           |              |                |             | 16s            | 651s        | 473s           | 2386s       |                |  |  |  |  |  |
| <b>Average Duration</b>                 |           |              |                |             |                |             |                |             |                |  |  |  |  |  |
| <i>Proprietary LVLMs</i>                |           |              |                |             |                |             |                |             |                |  |  |  |  |  |
| GPT-4V (OpenAI, 2023)                   | -         | -            | 43.7           | 49.2        |                | 60.7        | 53.5           | 59.9        | -              |  |  |  |  |  |
| GPT-4o (OpenAI, 2024)                   | -         | -            | 64.6           | 64.6        |                | 66.7        | 65.3           | 71.9        | 34.7           |  |  |  |  |  |
| Gemini-1.5-Pro (Team et al., 2024)      | -         | -            | 60.5           | 61.8        |                | 64.4        | 67.4           | 75.0        | 33.1           |  |  |  |  |  |
| <i>Open-Sourced LVLMs</i>               |           |              |                |             |                |             |                |             |                |  |  |  |  |  |
| InternVL2 (Chen et al., 2024b)          | 8B        | -            | 65.8           | 64.0        |                | 54.6        | -              | -           |                |  |  |  |  |  |
| Qwen2-VL (Wang et al., 2024b)           | 7B        | -            | 67.0           | -           |                | -           | -              | 63.3        | -              |  |  |  |  |  |
| Qwen2.5-VL (Bai et al., 2025)           | 7B        | 768          | -              | -           |                | -           | -              | 65.1        | 45.3           |  |  |  |  |  |
| LLaVA-OneVision (Li et al., 2024a)      | 7B        | -            | 56.7           | 64.7        |                | 56.3        | -              | 58.2        | -              |  |  |  |  |  |
| LLaVA-OneVision (Li et al., 2024a)      | 72B       | -            | 59.4           | 68.0        |                | 61.3        | -              | -           | 26.9           |  |  |  |  |  |
| <i>Long-Video LVLMs</i>                 |           |              |                |             |                |             |                |             |                |  |  |  |  |  |
| VideoLLaMA2 (Zhang et al., 2023b)       | 7B        | 72           | 54.6           | 48.5        |                | -           | 42.1           | 47.9        | -              |  |  |  |  |  |
| LongVA (Zhang et al., 2024c)            | 7B        | 128          | -              | 56.3        |                | -           | 46.2           | 52.6        | -              |  |  |  |  |  |
| LLaMA-VID (Li et al., 2023b)            | 7B        | 1 FPS        | 41.9           | 33.2        |                | -           | -              | -           | 23.9           |  |  |  |  |  |
| Oryx (Liu et al., 2024)                 | 7B        | 1 FPS        | 63.9           | 67.5        |                | 55.3        | 50.3           | 58.3        | -              |  |  |  |  |  |
| Oryx-1.5 (Liu et al., 2024)             | 7B        | 1 FPS        | 67.6           | 67.5        |                | 56.3        | 51.2           | 58.8        | -              |  |  |  |  |  |
| <b>LongVU</b> (Shen et al., 2024b)      | <b>7B</b> | <b>1 FPS</b> | <b>66.9</b>    | <b>65.4</b> |                | <b>59.5</b> | <b>52.4</b>    | <b>60.6</b> | -              |  |  |  |  |  |
| <i>Video Agents</i>                     |           |              |                |             |                |             |                |             |                |  |  |  |  |  |
| VideoAgent (GPT-4) (Fan et al., 2024)   | -         | 87           | -              | -           |                | -           | 49.0           | 56.0        | -              |  |  |  |  |  |
| VideoTree (GPT-4o) (Wang et al., 2025b) | -         | 98           | -              | -           |                | -           | 53.1           | -           | -              |  |  |  |  |  |
| <b>UniTime</b> (Li et al., 2025b)       | <b>7B</b> | <b>128</b>   | -              | <b>66.5</b> |                | <b>56.5</b> | -              | -           | -              |  |  |  |  |  |
| LLaVA-Video (Zhang et al., 2024f)       | 7B        | 80           | 57.7           | 64.4        |                | 58.3        | 52.4           | 63.4        | 41.3           |  |  |  |  |  |
| w/ <b>ZoomV</b>                         | 64+16     | 58.1         | ( <b>0.4</b> ) | 68.1        | ( <b>3.7</b> ) | 60.9        | ( <b>2.6</b> ) | 53.9        | ( <b>1.5</b> ) |  |  |  |  |  |
| InternVL2.5 (Chen et al., 2024b)        | 8B        | 80           | 70.1           | 67.1        |                | 60.6        | 52.2           | 63          | 41.8           |  |  |  |  |  |
| w/ <b>ZoomV</b>                         | 64+16     | 70.3         | ( <b>0.2</b> ) | 70.0        | ( <b>2.9</b> ) | 63.3        | ( <b>2.7</b> ) | 53.9        | ( <b>1.7</b> ) |  |  |  |  |  |
| Qwen2.5-VL (Team, 2025)                 | 7B        | 80           | 66.0           | 65.9        |                | 59.0        | 52.0           | 63.5        | 40.0           |  |  |  |  |  |
| w/ <b>ZoomV</b>                         | 64+16     | 66.0         | ( <b>0.0</b> ) | 67.0        | ( <b>1.1</b> ) | 61.0        | ( <b>2.0</b> ) | 53.6        | ( <b>1.6</b> ) |  |  |  |  |  |
| <b>InternVL3</b>                        | 8B        | 80           | 69.5           | 67.5        |                | 60.8        | 54.6           | 65.3        | 43.1           |  |  |  |  |  |
| w/ <b>ZoomV</b>                         | 64+16     | 69.5         | ( <b>0.0</b> ) | 68.5        | ( <b>1.0</b> ) | 63.6        | ( <b>2.8</b> ) | 54.9        | ( <b>0.3</b> ) |  |  |  |  |  |
| <b>Our ZoomV</b>                        | 73.6      | 52.4         | 24.5           | 48.6        | 61.0           | 43.0        | 26.1           | 43.9        | 48.4           |  |  |  |  |  |
|   |           |              |                |             |                |             |                |             | 36.4           |  |  |  |  |  |
|   |           |              |                |             |                |             |                |             | 36.7           |  |  |  |  |  |
|   |           |              |                |             |                |             |                |             | 76.5           |  |  |  |  |  |

407 **Table 2: Comparison of ZoomV with other LVLMs on video grounding results.** The results  
 408 include two temporal-sentence and one temporal-question grounding benchmarks.  
 409

| Model                                  | Charades-STA |             |             |             | ActivityNet-Captions |             |             |             | ReXTime      |              |              |             |
|--|--------------|-------------|-------------|-------------|----------------------|-------------|-------------|-------------|--------------|--------------|--------------|-------------|
|  | R@0.3        | R@0.5       | R@0.7       | mIoU        | R@0.3                | R@0.5       | R@0.7       | mIoU        | R@0.3        | R@0.5        | mIoU         | VQA         |
| CG-DETR (Moon et al., 2023)            | 70.4         | 58.4        | 36.3        | 50.1        | -                    | -           | -           | -           | 31.3         | 16.6         | 23.8         | -           |
| UniVTG (Lin et al., 2023)              | 72.6         | 60.2        | 38.6        | 52.1        | -                    | -           | -           | -           | 41.3         | 26.8         | 28.1         | -           |
| LITA (Huang et al., 2024b)             | -            | -           | -           | -           | -                    | -           | -           | -           | 29.49        | 16.29        | 21.49        | 34.44       |
| SeViLA (Yu et al., 2023)               | 27.0         | 15.0        | 5.8         | 18.3        | 31.6                 | 19.0        | 10.1        | 23.0        | -            | -            | -            | -           |
| Valley (Luo et al., 2023)              | 28.4         | 1.8         | 0.3         | 21.4        | 30.6                 | 13.7        | 8.1         | 21.9        | -            | -            | -            | -           |
| VideoChat2 (Li et al., 2024b)          | 38.0         | 14.3        | 3.8         | 24.6        | 40.8                 | 27.8        | 9.3         | 27.9        | -            | -            | -            | -           |
| Momenter (Qian et al., 2024)           | 42.6         | 26.6        | 11.6        | 28.5        | 42.9                 | 23.0        | 12.4        | 29.3        | -            | -            | -            | -           |
| VTimeLLM (Huang et al., 2024a)         | 51.0         | 27.5        | 11.4        | 31.2        | 44.0                 | 27.8        | 14.3        | 30.4        | 28.8         | 17.4         | 20.1         | 36.1        |
| TimeChat (Ren et al., 2024)            | 46.7         | 32.2        | 15.7        | -           | -                    | -           | -           | -           | 14.4         | 7.6          | 11.6         | 40.0        |
| HawkEye (Wang et al., 2024e)           | 50.6         | 31.4        | 14.5        | 33.7        | 49.1                 | 29.3        | 10.7        | 32.7        | -            | -            | -            | -           |
| GroundedVideo-LLM (Wang et al., 2024a) | 54.2         | 36.4        | 19.7        | 36.8        | 46.2                 | 30.3        | 19.0        | 36.1        | -            | -            | -            | -           |
| Qwen2.5-VL (Bai et al., 2025)          | -            | <b>24.2</b> | <b>11.1</b> | <b>29.0</b> | -                    | <b>15.8</b> | <b>7.5</b>  | <b>21.1</b> | -            | -            | -            | -           |
| VideoChat-R1 (Li et al., 2025a)        | -            | <b>70.6</b> | <b>47.2</b> | <b>59.9</b> | -                    | <b>33.3</b> | <b>16.7</b> | <b>35.5</b> | -            | -            | -            | -           |
| Time-R1 (Wang et al., 2025a)           | <b>78.1</b>  | <b>60.8</b> | <b>35.3</b> | -           | <b>58.6</b>          | <b>39.0</b> | <b>21.4</b> | -           | <b>31.81</b> | <b>18.46</b> | <b>22.48</b> | <b>72.1</b> |
| UniTime (Li et al., 2025b)             | -            | <b>59.1</b> | <b>31.9</b> | <b>52.2</b> | -                    | <b>22.8</b> | <b>14.1</b> | <b>27.3</b> | -            | -            | -            | -           |
| <b>Our ZoomV</b>                       | 73.6         | 52.4        | 24.5        | 48.6        | 61.0                 | 43.0        | 26.1        | 43.9        | 48.4         | 36.4         | 36.7         | 76.5        |

426 events and timelines without TemporaLink. Besides, TemporaLink provides consistent improvements  
 427 on the challenging ReXTime, which requires a strong ability to reason across time.  
 428

429 **Effectiveness of TemporaLight.** To validate our TemporaLight effectiveness, we employ the  
 430 LLaVA-Video model as the baseline, and equip it with our two types of reflections to watch the  
 431 difference. The results are reported in Table 3. We find that our TemporaLight enhances the base-



(a) Effectiveness of TemporaLink. (b) Robustness to video length via TemporaLink.

Figure 5: Ablation studies on temporal grounding and ultra-long video length.

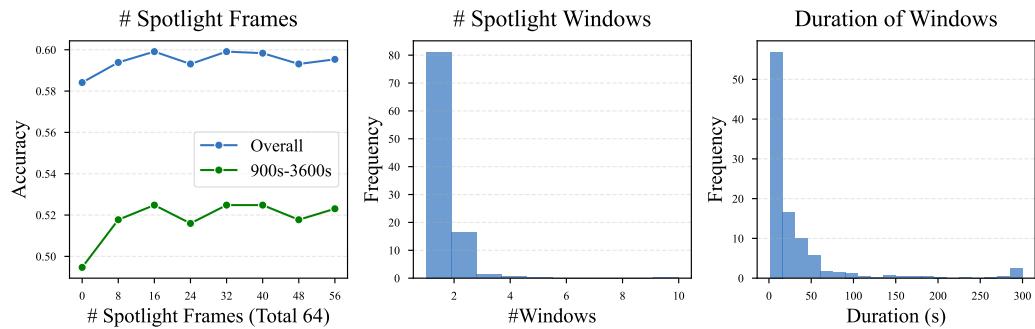


Figure 6: Impact of the spotlight frames on LongVideoBench.

line by 8.8% on LVBench, and 2.7% on LongVideoBench. Additionally, multiple-choice reflection, which offers more options for the model, shows better performance on video understanding tasks.

**Ablation Study of Frames Zoomed in.** We conduct experiments by varying the number of frames we zoom in, while keeping the frame budget fixed at 64. As shown in Figure 6 (left), introducing spotlight frames yields a significant boost in accuracy for both general cases and long videos. Our results suggest that 16 is an optimal setting, as it preserves global awareness while ensuring precise event retrieval. For a more in-depth understanding, we further analyze the number of spotlight windows and their duration distributions in Figure 6 (middle and right). The histogram of spotlight window counts reveals that most examples require only one or two spotlight windows, suggesting that many questions can be effectively answered with a small number of targeted events. Moreover, the spotlight duration histogram indicates that a majority of spotlighted events are relatively short (under 50 seconds). These findings highlight that a small number of well-chosen short spotlights is sufficient for significant improvements in long-video understanding, validating the effectiveness of our reflection-guided temporal search strategy in selecting relevant video moments efficiently.

**Robustness to Ultra-Long Video Lengths.** ZoomV shows noticeable improvements for video understanding models in Figure 5b. Specifically, for medium-length videos (*i.e.*, 180s–600s), simply applying the TemporaLink to supplement event details can yield consistent gains. Empirically, despite the significant loss of temporal dynamics in frame sampling, TemporaLink can still identify windows relevant to the questions based on limited visual cues. As the video length increases (*i.e.*, over 900s), it becomes increasingly challenging for TemporaLink to focus on useful events through sparse frames, and our search strategies are needed and result in significant improvements. Notably, for short videos, the framework maintains original performance as expected.

**Efficiency of ZoomV.** While we boost LVM performance via ZoomV, we uphold efficiency optimizations to ensure practicality. Initially, training a high-quality ZoomV on the LLaVA-Video model within 80 epochs only requires 8 hours utilizing 128 NVIDIA A100 GPUs. Furthermore, ZoomV introduces minimal additional latency during inference, and the runtime of per search step is 3483ms. We further optimize the multi-turn search by the prefix cache, as shown in Appendix Table 2. Eventually, we compare our method with other video-agent-style approach (*e.g.*, VideoTree).

486  
487  
488 Table 3: **Analysis of TemporaLight on video understanding tasks.**  
489  
490  
491

| Method                     | LVBench | LongVideoBench | LongVideoBench-Long |
|----------------------------|---------|----------------|---------------------|
| Baseline (LLaVA-Video)     | 41.3    | 58.3           | 48.4                |
| Yes/No Reflection          | 43.3    | 61.0           | 50.4                |
| Multiple Choice Reflection | 50.1    | 61.0           | 51.8                |

492 Table 4: **Detailed Comparison of ZoomV on efficiency and accuracy on EgoSchema.**  
493  
494

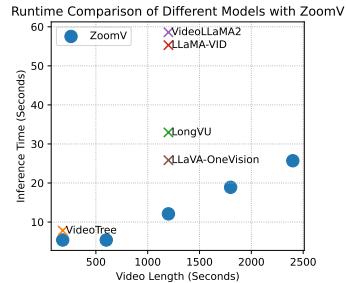
| Method    | grounding (s) | reflect (s) | caption (s) | keyfr. (s) | QA (s) | overall (s) | acc. (%) |
|-----------|---------------|-------------|-------------|------------|--------|-------------|----------|
| VideoTree | —             | —           | 1.6         | 4.4        | 1.8    | 7.8         | 63.6     |
| ZoomV     | 1.6           | 1.9         | —           | —          | 1.9    | 5.4         | 63.7     |

499 As shown in Table 4, while achieving higher accuracy, our method requires only 5.4s under the  
500 optimal parameters on a typical long-video dataset, compared to 7.8s for VideoTree, yielding an  
501 acceleration of approximately 30.8%.

502 To further validate the efficiency, we compare the inference time  
503 of ZoomV with other agents and long-video models across varying  
504 durations. As illustrated in Figure 7, we report the run-  
505 time of ZoomV in the worst-case scenario (*i.e.*, searching to the  
506 finest granularity), benchmarking it against state-of-the-art mod-  
507 els including VideoTree (Wang et al., 2025b), LongVU (Shen  
508 et al., 2024b), LLaVA-OneVision (Li et al., 2024a), VideoL-  
509 LaMA2 (Zhang et al., 2023b), and LLaMA-VID (Li et al.,  
510 2023b). ZoomV exhibits a linear scaling of inference time  
511 with video duration, yet remains highly efficient. Specifically,  
512 for short videos, ZoomV completes inference in just  $\sim$  5 sec-  
513 onds, outperforming VideoTree. Even as the duration extends to  
514 1200 seconds, ZoomV requires only  $\sim$  12.1 seconds. In sharp  
515 contrast, LongVU takes 33 seconds, while VideoLLaMA2 and  
516 LLaMA-VID incur significantly higher latency. These results  
517 underscore ZoomV’s superior scalability and efficiency for long-  
518 video understanding.

## 5 CONCLUSION

519 This paper introduces ZoomV, a novel framework for long-video understanding that emulates a  
520 human-like hierarchical temporal search. ZoomV proposes TemporaLink to retrieve key events  
521 and TemporaLight to verify predictions and guide the search direction. ZoomV achieves state-  
522 of-the-art performance across diverse video benchmarks, demonstrating significant gains in long-  
523 video QA and temporal grounding tasks. Furthermore, comprehensive ablation studies confirm the  
524 effectiveness of each component and underscore the importance of specialized designs for ultra-long  
525 video analysis. Finally, ZoomV bridges the gap between human cognitive strategies and model-  
526 based video analysis, providing a robust and interpretable solution for long video tasks.



527 Figure 7: **Runtime comparison**  
528 **across different video durations.** We report the worst-case  
529 scenario for ZoomV, demonstrating superior efficiency compared  
530 to fixed-cost long-video models.

540 REFERENCES  
541

542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
544 report. *arXiv preprint arXiv:2303.08774*, 2023.

545 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
546 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,  
547 2025.

548 Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A  
549 large-scale video benchmark for human activity understanding. In *CVPR*, pp. 961–970, 2015.

550 Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang.  
551 Rextime: A benchmark suite for reasoning-across-time in videos. In *NeurIPS*, 2024a.

552 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
553 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning  
554 for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer*  
555 *vision and pattern recognition*, pp. 24185–24198, 2024b.

556 Yue Fan, Xiaojian Ma, Ruijie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A  
557 memory-augmented multimodal agent for video understanding. In *European Conference on Com-  
558 puter Vision*. Springer, 2024.

559 Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu  
560 Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evalua-  
561 tion benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

562 Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via  
563 language query. In *ICCV*, pp. 5267–5275, 2017.

564 Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Xi Chen, and Bo Zhao. Vtg-llm: Integrating  
565 timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint*  
566 *arXiv:2405.13382*, 2024.

567 Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell.  
568 Localizing moments in video with temporal language. 2018.

569 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,  
570 and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

571 Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp  
572 video moments. In *CVPR*, 2024a.

573 De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding  
574 Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *arXiv preprint*  
575 *arXiv:2403.19046*, 2024b.

576 Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas  
577 Bertasius. Video recap: Recursive captioning of hour-long videos. In *CVPR*, 2024.

578 Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez,  
579 Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston,  
580 Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam  
581 Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion,  
582 Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei,  
583 Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared  
584 Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022.

585 Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural  
586 language queries. In *NeurIPS*, pp. 11846–11858, 2021.

594 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei  
 595 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint*  
 596 *arXiv:2408.03326*, 2024a.

597 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-  
 598 training for unified vision-language understanding and generation. In *International conference on*  
 599 *machine learning*. PMLR, 2022.

600 KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang,  
 601 and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*,  
 602 2023a.

603 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,  
 604 Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In  
 605 *CVPR*, 2024b.

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

609 Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language  
 610 models. *arXiv preprint arXiv:2311.17043*, 2023b.

611 Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Uni-  
 612 versal video temporal grounding with generative multi-modal large language models. *CoRR*,  
 613 *abs/2506.18883*, 2025b. doi: 10.48550/ARXIV.2506.18883. URL <https://doi.org/10.48550/arXiv.2506.18883>.

614 Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jin-  
 615 peng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal  
 616 grounding. In *ICCV*, 2023.

617 Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in  
 618 words. *Trans. Mach. Learn. Res.*, 2022, 2022.

619 Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-  
 620 demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*,  
 621 2024.

622 Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu,  
 623 Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced  
 624 ability. *arXiv preprint arXiv:2306.07207*, 2023.

625 Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang.  
 626 Mmg-vid: Maximizing marginal gains at segment-level and token-level for efficient video llms.  
 627 *arXiv preprint arXiv:2508.21044*, 2025.

628 WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-  
 629 dependency calibration in video representation learning for temporal grounding. *arXiv preprint*  
 630 *arXiv:2311.08835*, 2023.

631 OpenAI. GPT-4 technical report. *CoRR*, *abs/2303.08774*, 2023.

632 OpenAI. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, May 2024.

633 Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang  
 634 Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning.  
 635 In *ICML*, 2024.

636 Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal  
 637 large language model for long video understanding. In *CVPR*, 2024.

638 Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Ar-  
 639 belle. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint*  
 640 *arXiv:2404.00459*, 2024.

648 Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jian-  
 649 wei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through  
 650 tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024a.

651

652 Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun  
 653 Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J.  
 654 Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra.  
 655 Longvu: Spatiotemporal adaptive compression for long video-language understanding. *CoRR*,  
 656 abs/2410.17434, 2024b. doi: 10.48550/ARXIV.2410.17434. URL <https://doi.org/10.48550/arXiv.2410.17434>.

657

658 Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe  
 659 Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory  
 660 for long video understanding. In *CVPR*, 2024.

661 Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training  
 662 techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

663

664 John Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and*  
 665 *instruction*, 4(4):295–312, 1994.

666

667 Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe  
 668 sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision*  
 669 and *Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29118–  
 670 29128. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02711.  
 671 URL [https://openaccess.thecvf.com/content/CVPR2025/html/Tang\\_Adaptive\\_Keyframe\\_Sampling\\_for\\_Long\\_Video\\_Understanding\\_CVPR\\_2025\\_paper.html](https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html).

672

673 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,  
 674 Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal under-  
 675 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

676

677 Qwen Team. Qwen2.5-vl, January 2025. URL <https://qwenlm.github.io/blog/qwen2.5-vl/>.

678

679 Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng  
 680 Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video  
 681 large language models. *arXiv preprint arXiv:2410.03290*, 2024a.

682

683 Peng Wang, Shuai Bai, and et. al. Qwen2-vl: Enhancing vision-language model’s perception of the  
 684 world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

685

686 Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin  
 687 Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding  
 688 benchmark, 2024c.

689

690 Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video  
 691 understanding with large language model as agent. In *European Conference on Computer Vision*.  
 Springer, 2024d.

692

693 Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju,  
 694 Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin,  
 695 Jian Luan, and Qin Jin. Time-r1: Post-training large vision language model for temporal video  
 696 grounding. *arXiv preprint arXiv:2503.13377*, 2025a.

697

698 Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawk-  
 699 eye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*,  
 700 2024e.

701

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and  
 702 Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long  
 703 videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025b.

702 Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context  
 703 interleaved video-language understanding, 2024.

704

705 Penghao Wu and Saining Xie. V\*: Guided visual search as a core mechanism in multimodal llms.  
 706 *CVPR*, 2024.

707

708 Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded  
 709 video question answering. In *CVPR*, 2024.

710

711 Zikai Xie. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting  
 712 for large-language-models. *CoRR*, abs/2408.05093, 2024.

713

714 Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model  
 715 for video localization and question answering. In *NeurIPS*, 2023.

716

717 Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu,  
 718 Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning  
 719 to query frames for video large language models, 2025. URL <https://arxiv.org/abs/2410.03226>.

720

721 Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueling Zhuang, and Dacheng Tao. Activitynet-  
 722 qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

723

724 Jeffrey M Zacks and Khena M Swallow. Event segmentation. *Current directions in psychological  
 725 science*, 16(2):80–84, 2007.

726

727 Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas  
 728 Bertasius. A simple llm framework for long-range video question-answering, 2023a.

729

730 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language  
 731 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023b.

732

733 Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-  
 734 vstream: Memory-based real-time understanding for long video streams. 2024a.

735

736 Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal.  
 737 Generative verifiers: Reward modeling as next-token prediction. *CoRR*, abs/2408.15240, 2024b.

738

739 Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue  
 740 Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision.  
 741 *arXiv preprint arXiv:2406.16852*, 2024c.

742

743 Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang  
 744 Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in  
 745 mllms. *arXiv preprint arXiv:2506.10967*, 2025.

746

747 Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gu-  
 748 dovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsifi-  
 749 cation for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024d.

750

751 Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang,  
 752 Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. Unveiling the tapestry of consistency in  
 753 large vision-language models. *arXiv preprint arXiv:2405.14156*, 2024e.

754

755 Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video  
 756 instruction tuning with synthetic data, 2024f.

757

758 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,  
 759 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.  
 760 Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

761

762 Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang,  
 763 Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video  
 764 understanding. *arXiv preprint arXiv:2406.04264*, 2024.

756 **A THE USE OF LARGE LANGUAGE MODELS**  
757758 This paper makes use of an LLM solely for the purpose of polishing paragraph-level text.  
759760 **B ABSOLUTE TIMESTAMP CALIBRATION**  
761763 As stated in the main text, we utilize quantized integer timestamps to reduce the learning diffi-  
764 culty. However, the frame rate during frame extraction is often low for long videos, while man-  
765 ual annotations are done at a high frame rate. As a result, not every frame corresponding to  
766 a manually annotated time can be sampled. For example, the sampled frames are located at  
767 0s, 3s, ..., 67s, 70s, 73s, 77s, 80s, 83s, 86s, 89s and the target windows are serialized  
768 as [[72, 82], [84, 89]]. This problem introduces potential optimization challenges for text-  
769 oriented objectives. To address this, we propose the Absolute Temporal Calibration (ATC) method,  
770 which precisely aligns the annotated timestamps with the video decoding and frame extraction times.  
771 This calibration precisely aligns the annotated timestamps with the video’s specific frame time,  
772 thereby preventing the model from performing unnecessary frame interpolation during the learn-  
773 ing process. Specifically, in the example above, the target windows will first be adjusted to [[73,  
774 83], [83, 89]]. Subsequently, we will merge the overlapping windows caused by quantization  
775 errors, *i.e.*, calibrated target is [[73, 89]]. ATC ensures that the model can focus on temporal  
776 understanding without dealing with temporal discrepancies, thereby enhancing the model’s learning  
777 efficiency and temporal accuracy.778 **C INSTRUCTION TUNING**  
779780 The objective of Instruction Tuning is to equip the model with the ability to understand the Tempo-  
781 raLink.782 Table 1: Various tasks of our instruction dataset with the corresponding number of samples. {r}  
783 donate a list of time ranges corresponding to spotlighted video clips.  
784

| 785 <b>Tasks</b>     | 786 <b>Sources</b> | 787 <b>Instructions</b>  | 788 <b># of Samples</b> |
|----------------------|--------------------|--|-------------------------|
| 787 Spotlight        | QVHighlights       | Given the video and the query, find the relevant windows.                      | 7218                    |
|                      | Grounded-VideoLLM  | Provide the timestamps that correspond to the Answer.                          | 51918                   |
| 789 Reflection       | ReXTime            | Proposed time range: {r}. Is the proposed time range relevant to the question? | 19390                   |
|                      | Grounded-VideoLLM  | Proposed time range: {r}. Is the proposed time range relevant to the question? | 15220                   |
| 791 General Answer   | Grounded-VideoLLM  | General Video-QA instructions  | 107806                  |
|                      | LLaVA-Video        | General Video-QA instructions  | 79389                   |
|                      | NextQA             | Please respond with only the letter of the correct answer.                     | 6278                    |
| 793 Spotlight Answer | Moment-10M         | Please watch the clip of {r} and answer the question.                          | 42071                   |
|                      | Grounded-VideoLLM  | Please answer the question based on the detailed clip of {r}.                  | 17214                   |

795 **Datasets** As shown in Table 1 in the appendix, the training dataset is composed of four distinct  
796 tasks, all derived from existing open-source datasets. By introducing specialized instructions, we  
797 enhance the model’s capabilities in a cost-effective manner. The “Answering” capability is divided  
798 into two components: General Answering, which covers basic question-answering tasks like the  
799 most of LVLMs, and spotlighted answering, where answers are enriched using grounded video clips  
800 identified through a prior search for relevant spotlighted content.  
801802 **D EFFECTIVENESS AND EFFICIENCY TRADE-OFF IN SEARCH**  
803805 The reflection confidence threshold  $\epsilon$  and the minimum sub-event duration  $\Delta$  govern the search  
806 procedure. These hyperparameters jointly mediate the effectiveness-efficiency trade-off. From an  
807 effectiveness perspective, as validated in Figure 1, higher  $\epsilon$  and lower  $\Delta$  values improve accuracy at  
808 the cost of increased search steps. Reducing  $\Delta$  from 2400s to 600s with  $\epsilon = 0.8$  elevates LVbench  
809 accuracy from 46.8% to 49.5%, while finer-grained searches with  $\Delta = 300s$  do not result in im-  
provements. Regarding efficiency, the best-case complexity remains constant when  $\Delta$  exceeds the

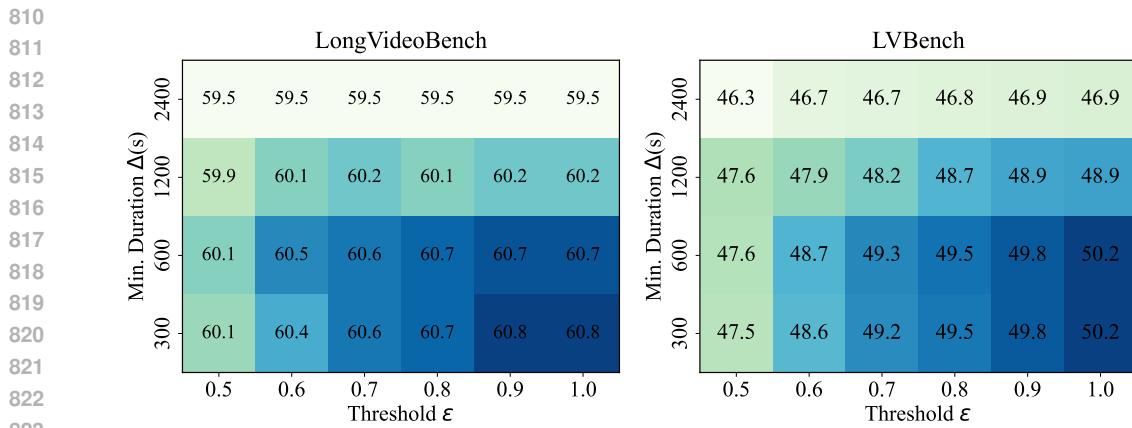


Figure 1: **Effectiveness and efficiency trade-off** with confidence threshold  $\epsilon$  and sub-event duration threshold  $\Delta$ .

Table 2: Optimization of ZoomV in Efficiency via prefix cache.

| Primitives                   | encoded frames | prefill (ms) | decode (ms) | overall (ms) |
|------------------------------|----------------|--------------|-------------|--------------|
| Grounding                    | 64             | 1157         | 424         | 1581         |
| Reflection                   | 80             | 1496         | 406         | 1902         |
| Reflection (w/ prefix cache) | 16             | 299.2        | 406         | 745          |

video length, while the worst-case complexity scales linearly. Specifically, when  $\Delta$  is larger than the video length, the search only executes a single step. In contrast, when  $\Delta$  is smaller than the video length, setting  $\epsilon = 1$  forces exhaustive traversal of all sub-events. The search prioritizes high-confidence segments through a priority queue, emulating human-like coarse-to-fine understanding. Empirical experiments demonstrate that  $\epsilon = 0.5$  requires only an average of 1.6 search steps while maintaining 99.5% of peak accuracy on LongVideoBench when  $\Delta = 1200s$ .

## E QUALITATIVE ANALYSIS

To further illustrate how **ZoomV** addresses challenges inherent in long-video understanding, we conduct a series of case studies on tasks involving temporal perception and chronological relations Wu et al. (2024). A core difficulty for LVLMS lies in insufficient temporal details, which often leads to misinterpretations of events. ZoomV mitigates this issue by integrating human-like *Spotlight* and *Reflection* mechanisms, allowing for more precise event retrieval.

For example, Figure 2 illustrates a case spanning 275 seconds, in which a man is sitting in front of a mirror. At the global (coarse) sampling level, only sparse frames can be observed, making it difficult to discern the subtle motion of his hands. The TemporalLink component in ZoomV addresses this issue by spotlighting a more fine-grained window from the 249th to the 275th second. Within this localized segment, the frame rate is increased, revealing that the man’s hands are clasped together—an action easily missed under low-frequency sampling. This example demonstrates how TemporalLink adaptively zooms in on the essential moments of a long video, capturing subtle actions that would otherwise be overlooked. Additionally, Figure 3 and Figure 4 showcase object attribute change and appearance order cases.

Figure 5 illustrates how ZoomV discerns sequential relationships between events in an ultra-long video through a hierarchical, coarse-to-fine search. In this example, ZoomV first identifies a large time window that roughly contains the relevant events. Upon noticing the disappearance of the white car, the search narrows to the 600-second sub-event window. Within this finer scope, ZoomV uses spotlight frames to focus on critical moments, identifying the appearance of a red car and a person. By progressively refining, ZoomV effectively captures the sequential flow of events, mimicking the way humans would search through long videos by zooming in on key events.

864  
865  
866  
867  
868

869 **Full Video Frames (275s)**



870 **Question:** In a room with a white background, there is a man with short hair wearing a short-sleeved T-shirt, sitting in front of a mirror.  
871 What is he doing at this moment?

872 **Options:** (A) Raised both hands upwards, (B) Shaking head, (C) Stood up, (D) Crying, (E) Hands clasped together

873 **Spotlight Frames with predicted windows [[249, 275]]**



874 **Figure 2: Illustration of the *subtle temporal dynamic challenge*.**

875  
876  
877  
878  
879  
880  
881  
882  
883

884 **Full Video Frames (186s)**



885 **Question:** Under the blue sky, a yellow car is driving on the road, with a truck behind it. There are tall trees planted in the greenbelt on both sides of the  
886 road. When a basketball hoop and a man in a black t-shirt appear around the yellow car, what change occurs with the yellow car?

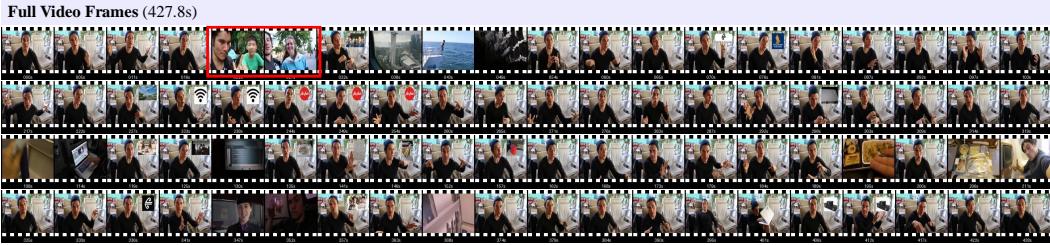
887 **Options:** (A) The door of the yellow car changes from closed to open. (B) The door of the yellow car shows some cartoon drawings. (C) The body of the  
888 yellow car gets sprayed with paint. (D) Some cartoon drawings appear on the body of the yellow car. (E) The headlights of the yellow car show some cartoon  
889 drawings.

890 **Spotlighted Videos with predicted windows [[121,123]]**



891 **Figure 3: Illustration of the *object attribute change challenge*.**

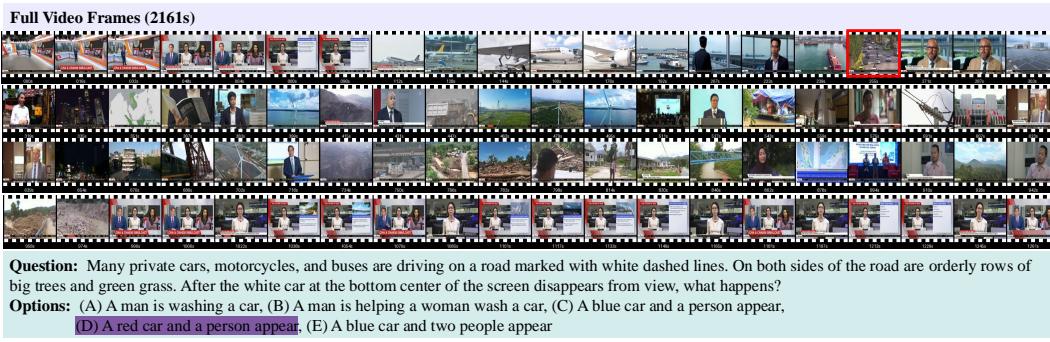
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903

918  
919  
920

922 **Question:** A man wearing a grey hat and black clothes is sitting on an off-white chair. The chair to his right is empty, and beside the empty chair, there is a potted plant. The wall behind him is blue. Among the photos that the man is showing, which photo appears first?

923 **Options:** (A) solo photo, (B) A group photo of four people, (C) A group photo of five people, (D) A group photo of two people, (E) A group photo of three people

924 **Spotlight Videos with multiple predicted windows [[0, 5], [16, 27]]**

938 **Figure 4: Illustration of the *object before/after object* challenge.**

940 **Question:** Many private cars, motorcycles, and buses are driving on a road marked with white dashed lines. On both sides of the road are orderly rows of big trees and green grass. After the white car at the bottom center of the screen disappears from view, what happens?

941 **Options:** (A) A man is washing a car, (B) A man is helping a woman wash a car, (C) A blue car and a person appear, (D) A red car and a person appear, (E) A blue car and two people appear

942 **TimeSearch Sub-event (600s)**



943 **Spotlight Frames (15s)**

964 **Figure 5: Illustration of the *event after event* challenge.**965  
966  
967