A Training-Free Video Moment Retrieval Framework via Selection from Multiple Visual Prompts

Xun Mo[1*] **Shengtao Guo**[1*] **Yongwei Nie**[1†]
Fei Ma[2] **Xuemiao Xu**[1] **Chengjiang Long**[3‡]
[1]South China University of Technology, Guangzhou, China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[3]Bytedance Inc., San Jose, CA 95110, USA
202421044335@mail.scut.edu.cn, nieyongwei@scut.edu.cn
https://github.com/MachineT-T/SelectVP

Abstract

Video Moment Retrieval (VMR) aims to localize a temporal segment in a video based on a natural language query. While Multi-modal Large Language Models (MLLMs) are increasingly adopted for VMR due to their strong video understanding capability, they still face several limitations: (1) imprecise perception of event boundaries, (2) rigid dependency on precise object identification, and (3) inability to handle temporarily interrupted events. To address these, we propose three visual prompts: Event Segmentation (ES) prompt, Object Detection (OD) prompt, and Keyframe Marking (KM) prompt. Each prompt improves performance when applied individually, yet direct fusion degrades results because mixing prompts overcomplicates video content and inaccurate labels introduce noise. Although certain visual prompts may mislead the model, at least one of the three is likely to yield a positive effect. Even in the worst case where all three perform poorly, the prediction can fall back to the baseline. Motivated by this, we introduce a selection mechanism based on Yes/No Probability-Difference score to choose the most reliable prediction among the results produced by the three visual prompts and the baseline. We evaluate our method in a training-free setting across three datasets and four models, and obtain state-of-the-art performance.

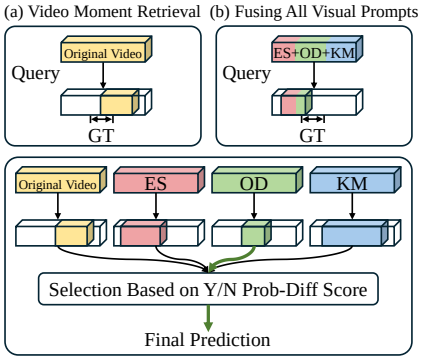
1 Introduction

Video Moment Retrieval (VMR) aims to localize temporal boundaries of specific events in untrimmed videos given natural language queries (as in Fig. 1 (a)). This task is fundamental to video understanding applications like video question answering and content-based retrieval.

Current VMR approaches follow three primary directions. The first trains specialized models from scratch [30, 49, 89, 20], typically extracting distinct features from video sequences and textual queries to fuse multi-modal features for timestamp prediction via meticulous architectures. The second category leverages pre-trained vision-language models (VLMs) to localize event intervals by computing frame-query similarity [91, 75, 35]. The third utilizes the video understanding of recent Multi-modal Large Language Models (MLLMs) to identify event timings through conversational interfaces. Given their extensive pre-training, these MLLM-based methods support supervised fine-tuning and reinforcement learning [64, 1, 82] or direct inference in training-free scenarios [91, 55, 5, 73, 86], significantly reducing requirements for computational resources and labeled data.

*Equal contribution.
†Corresponding author.
‡This work was done when Dr. Chengjiang Long was with Meta.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(c) Selection from Multiple Visual Prompts

	R1@0.3	R1@0.5	R1@0.7	mIoU
Baseline	71.0	48.7	22.6	45.9
ES	71.2	49.7	22.4	45.9
OD	71.5	49.9	23.1	46.6
KM	72.0	50.1	24.1	46.9
ES+OD	69.0	47.6	20.6	44.1
ES+KM	69.2	46.4	19.3	43.7
OD+KM	72.1	49.1	22.4	46.5
ES+OD+KM	67.8	45.9	19.1	43.1

(d) Comparison between Single Prompts and Fused Prompts

Figure 1: (a) Video Moment Retrieval aims to locate segments corresponding to a query in an untrimmed video. (b) Fusing multiple visual prompts into a single video may lead to a performance degradation. (c) Our method utilizes a Yes/No Probability-Difference score to select the optimal output among three visual prompts and baseline. (d) Although each of the three visual prompts independently improves performance, the fusion of these prompts results in a degradation.

Despite their robust video understanding, MLLMs exhibit critical limitations in VMR tasks. (1) These models demonstrate imprecise event boundary perception. For instance, given a query like "a person sits on a chair", the model often identifies the start time as the moment of contact, whereas the ground truth begins when the person initially approaches the chair. (2) MLLMs show a rigid dependency on precise object identification. For a query like "a person puts something in a bag", the model may misclassify the bag as a box and reject the correct temporal interval, despite high visual similarity. (3) Models tend to prematurely conclude events during temporary interruptions, failing to recognize action continuity when it resumes in subsequent frames.

To address these limitations, we propose three visual prompt types. The first is the Event Segmentation (ES) prompt, which clusters video features to partition events at cluster boundaries. By inserting title cards at these segmentation points, the model receives explicit cues regarding event transitions. The second is the Object Detection (OD) prompt, utilizing an open-vocabulary model to identify query-related objects and annotate them via bounding boxes. Adjusting detection confidence thresholds facilitates the identification of similar objects, mitigating rigid MLLM recognition constraints. The third is the Keyframe Marking (KM) prompt, which uses a VLM [56] to encode the query and frames, computing cosine similarity to highlight the top-k video frames as keyframes. This mechanism directs the model to focus on all event-related segments, even when temporally discontinuous.

As shown in Fig. 1 (d), all three types of visual prompts improve Qwen3-VL-8B [3] performance on Charades-STA [19]. However, simultaneously fusing multiple visual prompts into a video (e.g., Fig. 1 (b)) degrades performance compared to single prompts. Our analysis suggests that incorporating excessive visual prompts overcomplicates the video content. Furthermore, visual prompts from incorrect labels may introduce noise, adversely affecting model judgment. Although specific prompts might mislead the model, typically at least one yields a positive impact. Even in scenarios where all three prompts perform poorly, the system can revert to the original baseline result. Based on this intuition, we propose a selection mechanism (Fig. 1 (c)) to choose among the three prompt outcomes and the baseline. We calculate the probability-difference scores between "yes" and "no" tokens to measure reliability, selecting the candidate with the highest score as the final prediction.

To evaluate our method, we conduct experiments on Charades-STA [19], ActivityNet-Captions [34], and QVHighlights [36]. Results demonstrate that the proposed approach consistently enhances performance across multiple MLLMs, with the optimal configuration achieving state-of-the-art results on all three datasets. Furthermore, extensive ablation studies verify that the three visual prompts and selection mechanism play critical roles within the framework.

The contributions of this work are summarized as follows:

- We propose three distinct visual prompts to address the inherent limitations of applying MLLMs to the video moment retrieval (VMR) task, effectively enhancing their performance in a training-free setting.

- We introduce a Yes/No Probability-Difference score based selection mechanism, which integrates the three visual prompts into a robust framework capable of quantifying and selecting the most reliable prompt output.
- We evaluate our approach on standard VMR benchmarks and metrics, demonstrating its consistent effectiveness across various models and datasets and achieving SOTA results.

2 Related Work

Video Moment Retrieval with MLLMs. Early VMR approaches including proposal-based [74, 7, 84, 80, 87], proposal-free [81, 22, 8, 45, 85], reinforcement learning-based [27, 63, 69], and weakly-supervised approaches [48, 21, 15, 43, 9]. Despite the considerable progress made by these methods, they still face persistent challenges, such as bridging the semantic gap between visual and linguistic modalities, insufficient temporal context modeling, and limited generalization capabilities. Recently, Multi-modal Large Language Models (MLLMs) [32, 37, 11, 4] have demonstrated exceptional performance in the domain of video-language understanding. These models offer powerful cross-modal reasoning [12, 62, 50, 38, 28], superior temporal context modeling [26, 53, 60, 51, 29], and robust zero-shot generalization [78, 41, 67], significantly enhancing the potential for effective VMR. Consequently, many recent works have begun to leverage MLLMs for the VMR task [29, 57, 52, 24, 10, 23]. Some methods utilize MLLMs for boundary prediction and alignment [25, 66], while others employ them as high-level semantic localization tools [55, 54]. However, while these current methods claim to use MLLMs for VMR, they often only exploit them as auxiliary tools for tasks like text processing or video captioning, rather than for end-to-end reasoning. To truly harness the powerful video understanding capabilities of MLLMs, our work seeks to investigate how to address the limitations of MLLMs for the VMR task in an end-to-end manner based on varying visually prompted input videos.

Visual Prompts for MLLMs. Although textual prompting has been widely studied, it remains limited in describing fine-grained visual elements and detailed visual information [70]. Early works introduced visual prompts into pre-trained vision-language models [56, 39] using color-based co-referential markers [79] or red circles [58] to highlight image regions. With the rise of MLLMs [31, 61, 42], precise visual grounding has become increasingly important, as both users and models need to explicitly refer to visual components for accurate understanding. Consequently, visual prompts have been incorporated into MLLMs. Methods such as ViP-LLaVA [6], SoM [77], and Vigor [71] overlay simple shapes or masks on image regions and reference them via natural cues or numerical indices. Bounding boxes, commonly used as a visualization method in visual grounding, are also adopted as visual prompts [46, 16]. Inspired by the Chain-of-Thought (CoT) paradigm, visual prompts have further been used to represent intermediate reasoning steps [72, 93, 59, 86]. NumPro [73] notably improves VMR by simply inserting numerical indices into video frames, showing that even simple visual cues can significantly boost performance. Motivated by this, we aim to explore the potential of visual prompts to improve performance of VMR in training-free setting.

3 Method

3.1 Overview

Given an untrimmed video $V = \{v_i\}_{i=1}^{N}$ consisting of N frames and a textual query Q, MLLM-based Video Moment Retrieval (VMR) aims to localize the temporal segment corresponding to the queried event. Formally, this task is defined as $T = \mathcal{M}(V, Q, I)$, where $\mathcal{M}$ denotes the MLLM, I represents a task-specific instruction, and $T = [t^s, t^e]$ signifies the predicted temporal interval, with t^s and t^e being the start and end timestamps of the target event.

As illustrated in Fig. 2, the proposed framework operates through a multi-branch, training-free pipeline designed to dynamically select the most reliable result from a set of visual prompts. The process initiates by processing the original video through several parallel branches. Specifically, the framework maintains one branch for the original video and three additional branches in which the video is prompted with distinct visual prompts.

Each of these visually prompted videos is subsequently fed into a shared base MLLM to generate an initial set of candidate temporal segments. In the following step, the video portions corresponding

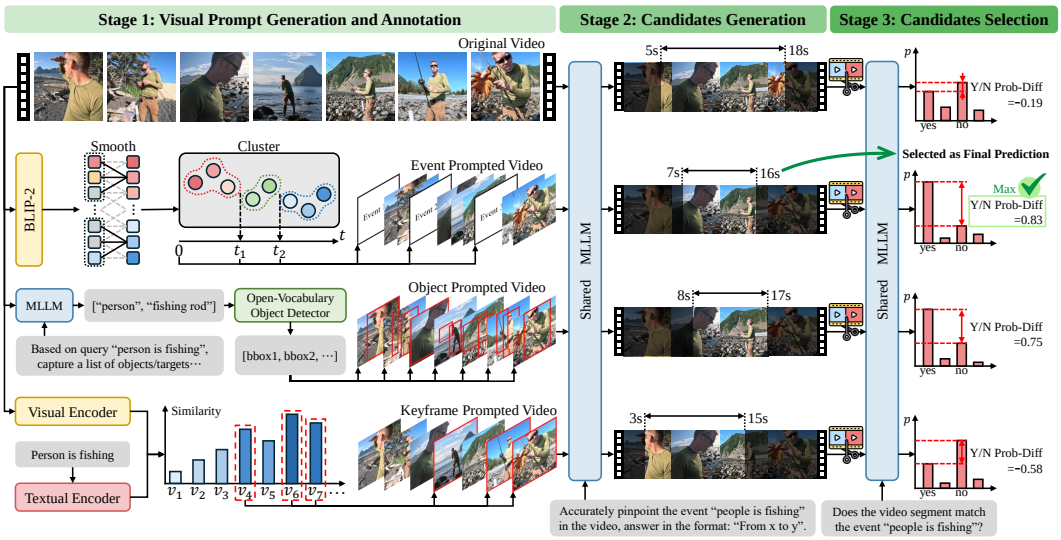


Figure 2: **Overview of our proposed training-free framework for video moment retrieval.** The pipeline consists of three stages: (1) Generation and Annotation of Visual Prompts, (2) Generation of Candidate VMR Results, and (3) Selection Based on Yes/No Probability-Difference Score.

to each candidate segment are cropped. Each cropped segment is then presented back to the same shared base model with a verification prompt, which requires the model to determine whether the video segment accurately corresponds to the given query and to provide an answer of "yes" or "no", represented as probability bars in the figure. Finally, the candidate segment associated with the highest Yes/No Probability-Difference score is selected as the final prediction for the input query.

3.2 Generation and Annotation of Visual Prompts

To address the limitations of MLLMs in VMR task, we propose three types of visual prompts from the perspectives of events, objects, and keyframes, namely Event Segmentation (ES) prompt, Object Detection (OD) prompt, and Keyframe Marking (KM) prompt. The initial stage of our framework involves the systematic generation and annotation of these prompts directly onto the video frames.

Event Prompting. The first type of visual prompt is the ES prompt. As shown in Fig. 2, we segment a video into multiple event clips and insert a title card at each clip's start. This addresses inaccurate MLLM boundary perception. For instance, in the event "a person sits on a chair", models often assume the event starts upon contact, though the ground truth begins when the person moves toward the chair. The ES prompt explicitly provides these boundaries, enabling more precise localization.

We adopt a clustering-based segmentation approach. First, a visual encoder [39] extracts frame features, followed by a temporal sliding window for smoothing. This ensures early-stage features (e.g., "a person moves toward a chair") encode subsequent action information (e.g., "the person sits on the chair"). We then cluster smoothed features using the MeanShift algorithm [14], which automatically determines cluster counts. We treat time t as a new event start whenever the cluster label at t differs from $t-1$, segmenting the video into distinct events.

Inspired by cinematic transitions, we insert a title card (a white frame labeled "Event i") at the beginning of each event. Letting b_i denote the ith card, the event-prompted video is:

$$V_{\text{event}} = [b_1, v_1, v_2, \ldots, b_i, \ldots, v_N]. \tag{1}$$

Object Prompting. The second visual prompt is the OD prompt. As shown in Fig. 2, we employ an open-vocabulary detection model to identify query-mentioned objects, annotating video frames with red bounding boxes. This addresses the MLLMs' rigid dependency on precise object identification. For instance, in "a person puts something in a bag", the model might misidentify the bag as a box and reject the correct interval, despite visual similarity. Visualizing detection results directs model

attention toward core query-related objects. Furthermore, adjusting the detection confidence threshold mitigates the rigid classification constraints inherent in MLLMs.

We first utilize an MLLM to extract a list of salient nouns $\mathcal{R}_{\text{o}}$ from query Q:

$$\mathcal{R}_{\text{o}} = \mathcal{M}(V, Q, I_o), \tag{2}$$

where I_{o} represents the instruction, specifically indicated as "*Based on the event description 'Q', your task is to capture a list of objects/targets that are within the description. These targets will be used for object detection and grounding via an object detection model. Please response in the form of a Python string list (e.g., ['object1', 'object2']).*" Subsequently, we use the open-vocabulary detection model MM-Grounding-DINO [88] to detect all nouns in $\mathcal{R}_{\text{o}}$ based on a confidence threshold σ. The detector $\mathcal{D}_\sigma$ identifies the bounding boxes for these objects and annotates them onto the original video to produce the object prompted video:

$$V_{\text{object}} = \mathcal{D}_\sigma(V, \mathcal{R}_{\text{o}}). \tag{3}$$

Keyframe Prompting. The third visual prompt is the KM prompt, which highlights the video frames most relevant to the query (see Fig. 2). When an event is temporarily interrupted, MLLMs often incorrectly perceive the event as concluded, thereby ignoring subsequent actions that continue after the pause. The KM prompt directs the model's attention to all segments closely associated with the query, even if they are temporally discontinuous, which addresses the issue of premature prediction termination.

We utilize the text and vision encoders provided by a VLM (e.g., CLIP [56]) to extract features from the query and video, respectively. The cosine similarity between these features is calculated, where higher similarity indicates a stronger correlation between a video frame and the query. We select the top-k frames with the highest similarity scores as keyframes and mark them in the video using red bounding boxes. The resulting keyframe prompted video can be represented as:

$$V_{\text{keyframe}} = \mathcal{C}_k(V, Q), \tag{4}$$

where $\mathcal{C}$ denotes the keyframe selection and marking process.

3.3 Generation of Candidate VMR Results.

We obtain four video variants: the original video V and three visually prompted versions V_{event}, V_{object}, and V_{keyframe}. By feeding each video separately into the MLLM $\mathcal{M}$ along with the query Q and instruction I, we generate a candidate temporal segment for each version:

$$\begin{aligned} T_{\text{origin}} &= \mathcal{M}(V, Q, I), \\ T_{\text{event}} &= \mathcal{M}(V_{\text{event}}, Q, I), \\ T_{\text{object}} &= \mathcal{M}(V_{\text{object}}, Q, I), \\ T_{\text{keyframe}} &= \mathcal{M}(V_{\text{keyframe}}, Q, I). \end{aligned} \tag{5}$$

Here, the instruction I is defined as "*Accurately pinpoint the event 'Q' in the video, answer in the format: 'From x to y'.*"

3.4 Selection Based on Yes/No Probability-Difference Score

As discussed in Section 1 and illustrated in Fig. 1 (d), fusing multiple visual prompts can overcomplicate the video content, while incorrect labels within these prompts may mislead the model. However, at least one of the three visual prompts typically exerts a positive influence on the model. Even in scenarios where all three prompts perform poorly, the system can degrade gracefully to the baseline performance. Guided by this intuition, we propose a selection mechanism capable of identifying the most reliable result from four candidates. To quantify the reliability of each candidate, we introduce the Yes/No Probability-Difference score.

For each candidate temporal interval $T_x = [t_x^s, t_x^e]$, where $x \in \{\text{origin}, \text{event}, \text{object}, \text{keyframe}\}$, we extract the corresponding video clip from the original video. This clip is then fed into the MLLM alongside a verification instruction I_{verify}: "*Does this video segment match the event 'Q'? You must*

Table 1: **Comparison with state-of-the-art zero-shot methods on three VMR benchmarks.** We report R1@m (for $m \in \{0.3, 0.5, 0.7\}$) and mIoU. Our framework, when applied to four different base MLLMs (+ours), consistently improves their performance. Notably, the application of the proposed method to Qwen3-VL-8B achieves state-of-the-art results among all approaches. The best results are indicated in **bold**, and the second-best results are underlined.

Method	Charades-STA				ActivityNet-Captions				QVHighlights			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
ChatVTG [55]	52.7	33.0	15.9	34.9	40.7	22.5	9.4	27.2	-	-	-	-
Moment-GPT [75]	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8	-	58.9	38.6	-
NumPro [73]	60.7	36.8	15.9	38.5	44.2	26.4	14.4	31.3	-	-	-	-
TimeSuite [83]	69.9	48.7	24.0	-	-	-	-	-	-	-	-	-
TFVTG [91]	67.0	50.0	24.3	44.5	49.3	27.0	13.4	34.1	-	-	-	-
VideoChat-R1.5-7B [76]	-	-	-	-	52.4	32.3	16.8	35.5	71.4	55.8	38.4	52.7
Chrono-GPT [47]	-	28.8	11.0	-	-	31.0	17.5	-	-	61.7	41.8	-
Llava-OneVision-1.5-8B [2]	31.6	9.0	1.9	22.1	31.7	17.8	10.5	23.9	37.7	23.0	13.0	27.3
+ours	36.9(+5.3)	12.3(+3.3)	3.0(+1.1)	25.0(+2.9)	33.5(+1.8)	18.9(+0.9)	11.0(+0.5)	25.1(+1.2)	39.4(+1.7)	24.2(+1.2)	13.4(+0.4)	29.3(+2.0)
InternVideo2.5-8B [65]	49.9	24.5	7.9	30.6	22.0	9.7	3.7	16.6	31.7	15.9	6.3	23.8
+ours	52.4(+2.5)	26.1(+1.6)	9.1(+1.2)	32.2(+1.6)	23.8(+1.8)	10.9(+1.2)	4.0(+0.3)	17.7(+1.1)	36.2(+4.5)	19.4(+3.5)	7.2(+0.9)	26.2(+2.4)
Qwen2-VL-7B [61]	61.2	37.4	16.9	39.1	44.5	26.8	14.9	31.6	56.5	34.8	18.4	39.0
+ours	67.2(+6.0)	42.8(+5.4)	18.9(+2.0)	42.6(+3.5)	47.1(+2.6)	28.6(+1.8)	15.6(+0.7)	33.3(+1.7)	63.0(+6.5)	41.0(+6.2)	21.4(+3.0)	43.6(+4.6)
Qwen3-VL-8B [3]	71.0	48.7	22.6	45.9	51.1	34.6	**20.9**	37.3	71.2	58.3	39.4	53.5
+ours	**73.8**(+2.8)	**53.0**(+4.3)	**25.2**(+2.6)	**48.2**(+2.3)	**53.2**(+2.1)	**35.7**(+1.1)	**20.9**	**38.3**(+1.0)	**76.3**(+5.1)	**61.9**(+3.6)	**42.2**(+2.8)	**57.0**(+3.5)

only answer 'yes' or 'no'." By performing a single forward pass, we obtain the probability distribution for the next token. The Yes/No Probability-Difference (Y/N Prob-Diff) score is defined as:

$$S_{\text{pd}}(T_x) = p(\text{“yes”}) - p(\text{“no”}), \tag{6}$$

where $p(\text{“yes”})$ and $p(\text{“no”})$ denote the probabilities that the next token is any case-variant of “yes” or “no”, respectively. Since the instructions explicitly require the model to output “yes” or “no”, the probabilities of other tokens are negligible and are disregarded. A higher score indicates greater confidence from the MLLM that the video clip aligns with the query. Consequently, we select the candidate with the highest Y/N Prob-Diff score as the final prediction:

$$\hat{T} = \arg\max_{T_x} S_{\text{pd}}(T_x). \tag{7}$$

4 Experiment

4.1 Experimental Setup

To demonstrate effectiveness, we evaluate our method on two standard VMR datasets: Charades-STA [19] and ActivityNet-Captions [34]. Following prior works [29, 52, 57, 66, 73], we adopt standard metrics: mean Intersection over Union (mIoU) and Recall@1 at various IoU thresholds m (R1@m), where $m \in \{0.3, 0.5, 0.7\}$. To verify generalization, we conduct additional experiments on QVHighlights [36]. Though originally for highlight detection, we adapt it for VMR following [75, 47, 76], employing the same metrics as our primary experiments.

4.2 Implementation Details

Following prior work [73, 75], we set the video frame rate to 1 FPS and resize frames to 336×336. We evaluate our method using representative open-source models: Qwen2-VL-7B [61], Qwen3-VL-8B [3], LLaVA-OneVision-1.5-8B [2], and InternVideo2.5-8B [65]. To ensure reproducibility, `do_sample` is set to `False`. For the ES prompt, we adopt the MeanShift algorithm [14] for clustering. For the OD prompt, we utilize MM-Grounding-DINO [88] with a confidence threshold $\sigma = 0.55$. For the KM prompt, we employ CLIP [56] for keyframe selection, retaining the top-5 frames with the highest similarity scores. All experiments are conducted on one NVIDIA A100 (40GB) GPU.

4.3 Main Results

Comparison with State-of-the-Art Methods. Table 1 presents a comprehensive comparison between our method and state-of-the-art (SOTA) approaches under the **zero-shot** setting. When applied to the Qwen3-VL-8B model, our method outperforms all zero-shot MLLM-based methods and achieves significant improvements over previous SOTA results. Specifically, on the Charades-STA dataset, our approach yields gains of 3.9%, 3.0%, 0.9%, and 3.7% in terms of R1@0.3, R1@0.5,

Table 2: **Visual prompt vs. textual prompt.** For all three prompt types, the textual prompt results in significantly lower performance across all metrics compared to the visual prompt.

		Charades-STA			
		R1@0.3	R1@0.5	R1@0.7	mIoU
ES	Visual	71.2	49.7	22.4	45.9
	Textual	64.7	42.7	20.1	42.1
OD	Visual	71.5	49.9	23.1	46.6
	Textual	61.9	35.4	16.5	39.1
KM	Visual	72.0	50.1	24.1	46.9
	Textual	60.7	37.3	15.2	38.8

Table 3: **Ablation study on different visual prompt combinations.**

BL	ES	OD	KM	Charades-STA R1@0.3	R1@0.5	R1@0.7	mIoU
✓				71.0	48.7	22.6	45.9
	✓			71.2	49.7	22.4	45.9
		✓		71.5	49.9	23.1	46.6
			✓	72.0	50.1	24.1	46.9
✓	✓			73.2	52.0	23.6	47.5
✓		✓		71.9	50.5	23.7	46.8
✓			✓	72.5	51.2	24.7	47.4
✓	✓	✓		73.5	52.7	24.5	47.9
✓	✓		✓	73.7	52.6	25.1	48.1
✓		✓	✓	72.7	51.8	24.7	47.6
✓	✓	✓	✓	73.8	53.0	25.2	48.2

R1@0.7, and mIoU, respectively. On the ActivityNet-Captions dataset, our method surpasses VideoChat-R1.5-7B [76] by 0.8%, 3.4%, 4.1%, and 2.8% across the same metrics. Furthermore, on the QVHighlights dataset, our method improves upon the previous SOTA approach, Chrono-GPT [47], by 0.2% and 0.4% in R1@0.5 and R1@0.7, respectively.

Effectiveness of Our Method Across MLLMs. We further evaluate the effectiveness and generalization capability of the proposed method by applying it to four representative models: LLaVA-OneVision-1.5-8B, InternVideo2.5-8B, Qwen2-VL-7B, and Qwen3-VL-8B. The performance of our approach is compared with the corresponding baselines for each model. As demonstrated in Table 1, the proposed method yields consistent improvements across all models and datasets, with gains in mIoU ranging from +1.0 to +4.6. Notably, the improvements on Qwen2-VL-7B are the most pronounced, as evidenced by the largest performance gains across the three evaluated datasets.

4.4 Ablation Studies

Visual Prompt vs. Textual Prompt. To provide the MLLM with supplementary information regarding events, objects, and keyframes, we employ a visual prompting strategy. We evaluate the effectiveness of visual prompts relative to textual prompts by conducting comparative experiments for event, object, and keyframe prompting using both modalities. Specifically, the textual prompts are formatted as descriptive strings (detailed in Section E), while the visual prompts follow our proposed method. As shown in Table 2, textual prompts consistently underperform visual prompts across all evaluation metrics. This performance disparity likely stems from the inherent modality gap between text and video. Specifically, the model struggles to align temporal timestamps and spatial coordinates mentioned in text with the raw video content.

Ablation Study on Different Combinations of Visual Prompts. We evaluate the performance of various combinations of the baseline and different visual prompts on the Charades-STA dataset, employing our proposed selection mechanism to determine the final prediction. As reported in Table 3, any single visual prompt independently outperforms the baseline, and integrating these visual prompts with the baseline yields further performance gains. Furthermore, all four evaluation metrics exhibit a consistent upward trend as the diversity of integrated visual prompts increases.

Is the Gain from Visual Prompts or from Mere Diversity? The preceding ablation study demonstrates that all evaluation metrics improve as the diversity of integrated visual prompts increases. This observation raises a critical question: *whether the performance gains stem solely from the diversity of the outputs rather than the specific contributions of the visual prompts.* To investigate this, we adjust the generation configuration (`do_sample=True`, `temperature=0.8`, `top_p=0.95`) to produce four diverse baseline outputs. We then compare the results of these multiple baseline candidates against multiple visual prompting candidates, maintaining an identical number of candidates for both groups. As illustrated in Fig. 3, the mIoU of the Multiple Visual Prompts consistently exceeds the mIoU of the Multiple Baseline as the number of candidates increases. These results indicate that well-designed visual prompts are essential for performance enhancement, rather than mere output diversity.

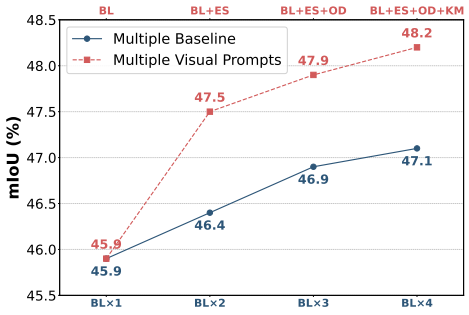


Figure 3: **Comparison between multiple baseline candidates and multiple visual prompting candidates on Charades-STA dataset.** "Multiple Baseline" (blue) generates N (where $N \in \{1, 2, 3, 4\}$) candidates using stochastic sampling (i.e., `do_sample=True`, `temperature=0.8`, `top_p=0.95`) by the original video. "Multiple Visual Prompting" (orange) generates N candidates using the proposed visual prompting methods.

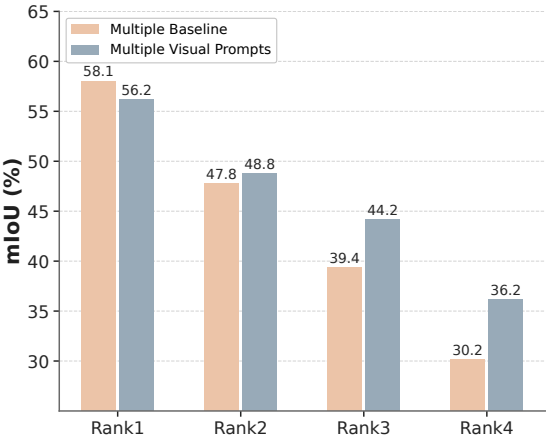


Figure 4: **Rank-wise comparison of candidate mIoU between Multiple Baseline and Multiple Visual Prompts on Charades-STA.** Multiple Visual Prompts exhibit a smaller performance spread across ranks, improving robustness to imperfect candidate selection.

Why Do Multiple Visual Prompts Outperform Multiple Baseline? To further understand this advantage, we rank the four candidates for each query by their IoU and compute the mIoU at each rank. As shown in Fig. 4, although Multiple Baseline achieves a higher best-case mIoU (58.1 vs. 56.2), Multiple Visual Prompts consistently outperform it at the remaining three ranks, with particularly large gains at Rank 3 and Rank 4. Notably, the gap between the best and worst ranks decreases from 27.9 to 20.0 points. This suggests that visual prompts not only introduce diversity but also reduce the variability in candidate quality, avoiding extremely poor predictions. Since the selection mechanism is not always accurate, such stability mitigates the performance degradation caused by incorrect selection, ultimately leading to better overall performance.

Ablation Study on Visual Prompt Generation Methods. In Table 4, we conduct a comprehensive ablation study on the methods used to generate visual prompts. Each branch is combined with baseline and evaluated with multiple variant methods to verify the influence of the underlying models and algorithms.

For the **ES branch**, we compare different clustering strategies. Compared to the random baseline (45.1 at R1@0.5), advanced clustering methods such as K-Means ($K = 3$) and MeanShift significantly improve performance, achieving 50.2 and 52.0 at R1@0.5, respectively. This indicates that more accurate segmentation of event boundaries benefits the overall framework.

For the **OD branch**, we test several object detection backbones. We observe that more powerful detectors yield better results: while simple detectors such as YOLO11 achieve moderate gains (49.2 at R1@0.5), advanced multi-modal models like MM-Grounding-DINO push the performance further, reaching 50.5 at R1@0.5, which also leads to the highest mIoU (46.8) in this

Table 4: **Comparison between different visual prompt generation method.** We ablate the generation methods for each category of visual prompt. Each type of visual prompt is independently combined with the baseline, followed by selection using our proposed visual-prompt selector. For all categories of visual prompts, performance is strongly correlated with the quality of the chosen method.

Variant Method		Charades-STA			
		R1@0.3	R1@0.5	R1@0.7	mIoU
Baseline		71.0	48.7	22.6	45.9
ES	+Random	69.1	45.1	19.7	43.9
	+EfficientGEBD [92]	72.4	50.2	22.4	46.5
	+K-Means [68]	72.7	50.2	22.3	46.5
	+MeanShift [14]	**73.2**	**52.0**	**23.6**	**47.5**
OD	+Random	71.2	48.8	22.7	46.1
	+RT-DETR [90]	71.6	49.3	23.2	46.2
	+YOLO11 [33]	71.4	49.2	22.9	46.2
	+YOLO-World [13]	71.5	49.8	23.5	46.5
	+LLMDet [18]	**71.9**	50.0	23.4	46.7
	+MM-Grounding-DINO [88]	**71.9**	**50.5**	**23.7**	**46.8**
KM	+Uniform	71.4	49.0	22.4	46.1
	+Random	71.2	49.4	22.9	46.3
	+MLLM [3]	72.4	50.9	24.1	47.2
	+CLIP [56]	**72.5**	**51.2**	**24.7**	**47.4**

branch. This trend demonstrates that the precision of visual object cues directly influences the downstream temporal reasoning.

For the **KM branch**, we explore different keyframe selection strategies. Compared with uniform and random sampling, leveraging large-scale pre-trained models (e.g., MLLM and CLIP) consistently improves results. In particular, CLIP-based keyframe selection achieves the best overall performance with 51.2 at R1@0.5 and 47.4 mIoU, highlighting the advantage of semantically meaningful frame selection.

Overall, the results show that the framework's performance depends on the quality of the adopted methods for generating visual prompts. As the accuracy of the visual representations improve, the final retrieval and localization performance is improved too. The proposed framework could benefit from future advances in these detection tasks.

Ablation Study on Candidate Selection Methods. To demonstrate the superiority of our proposed selection mechanism, we conduct an ablation study comparing four distinct selection strategies: "Random", "MLLM", "Avg. Log-Prob", and "Y/N Prob-Diff". As reported in Table 5, our "Y/N Prob-Diff" method achieves the highest performance across all metrics. Specifically, "Random" serves as a baseline by selecting a single candidate at random from the pool. The "MLLM" strategy involves feeding all candidates simultaneously into the model and querying it to identify the optimal choice. While this approach is straightforward, it yields suboptimal results. The "Avg. Log-Prob" strategy calculates the average log-probability of all tokens in the model output as a quality metric. However, because this metric incorporates every generated token, its effectiveness remains limited. In contrast, our proposed "Y/N Prob-Diff" method focuses on the binary probability difference of a single critical token, which more accurately reflects the confidence of the MLLM regarding the candidate quality, thereby achieving the best overall performance.

Table 5: **Comparison between different candidate selection methods.** "Random", "MLLM" (prompting MLLM to select), "Avg.Log-Prob" (average log-probability of all tokens), and our proposed "Y/N Prob-Diff" (Yes/No Probability-Difference score).

Method	Charades-STA			
	R1@0.3	R1@0.5	R1@0.7	mIoU
Random	69.7	47.4	21.3	44.8
MLLM	71.6	49.5	22.5	46.1
Avg. Log-Prob	71.5	49.4	23.0	46.3
Y/N Prob-Diff	**73.8**	**53.0**	**25.2**	**48.2**

4.5 Analysis of Selection Bias

To investigate whether the proposed selection mechanism exhibits any inherent bias during selection, we analyze the distribution of chosen visual prompts across the Charades-STA, ActivityNet-Captions, and QVHighlights datasets. As illustrated in Fig. 5, the selection distributions remain largely consistent across all three datasets. Apart from a marginally higher selection frequency for the ES prompt, no significant preference for any specific candidate is observed. This balanced distribution indicates that each of the four branches contributes significantly to the overall framework.

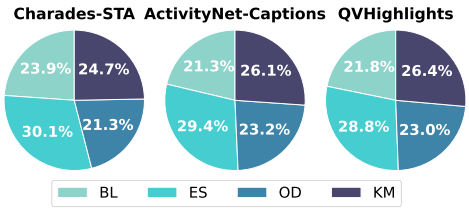


Figure 5: **Distribution of chosen visual prompts.** The distribution of chosen visual prompts remains largely consistent across three datasets. Apart from a marginally higher frequency for the ES prompt, the results show no pronounced selection bias.

4.6 Analysis of Computational Cost

Our approach comprises four independent branches, allowing for either serial execution to minimize memory overhead or parallel execution to reduce inference time. This modular design enables a trade-off between efficiency and performance by adjusting the number of active branches.

Inference Time. Although each branch involves inference and verification steps, the overall latency remains competitive. The inference step generates VMR results directly without extensive chains of thought; the verification step requires only a single forward pass rather than full autoregressive generation, ensuring its cost is substantially lower than the inference step. As shown in Table 6,

evaluations on ActivityNet-Captions demonstrate that our method achieves the highest mIoU with the lowest latency among multiple-inference approaches. Under serial execution, the framework maintains superior efficiency. With sufficient GPU memory, parallel execution even reduces latency to 2.4 s, comparable to the 2.2 s of the Qwen3-VL-8B baseline.

GPU Memory. Beyond the inherent fixed overhead of MLLMs, GPU memory consumption is primarily determined by token count. Since visual prompts are annotated onto frames, they do not increase tokens, except for a marginal increase from the insertion of prompt frames for ES prompt. Consequently, our serial implementation's memory overhead aligns with the Qwen3-VL-8B baseline, requiring approximately 20GB on Charades-STA and 32GB on ActivityNet-Captions. In parallel configurations, GPU memory scales linearly with the number of branches.

Table 6: **Comparison of inference time and mIoU across different inference types on ActivityNet-Captions dataset.** Compared to other multiple-inference methods, our approach achieves the superior performance with the shortest total latency under serial inference. With sufficient GPU memory, the inference time can be further reduced through parallel inference.

	Inference Type	Latency	mIoU
VideoChat-R1.5-7B [76]	Iterative Inference	11.0 s	16.8
VideoMind [44]	Agent	18.2 s	33.3
Ours	Serial Inference	9.6 s	**38.3**
Ours	Parallel Inference	**2.4 s**	**38.3**

5 Conclusion

This paper proposes a training-free VMR framework. To overcome challenges in applying MLLMs to VMR, we introduce three types of visual prompts. We observe that integrating multiple prompts simultaneously can overcomplicate video content and introduce noise through inaccurate cues. Intuitively, at least one proposed prompt contributes positively. Even if all perform poorly, the system can revert to the baseline result. Consequently, we employ a selection mechanism based on Yes/No Probability Difference scores to evaluate output reliability, choosing the most dependable result from the prompts and baseline as the final prediction. Extensive experiments on VMR benchmarks demonstrate state-of-the-art performance in a training-free setting. We further verify framework generalization across multiple open-source models. Ablation studies confirm that each prompt and the selection mechanism significantly contribute to the framework. Finally, analyses of selection bias and computational overhead illustrate that our selector maintains no significant preference and additional costs remain acceptable.

Acknowledgment

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2025A1515011884, 2026A1515010184), the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-24-17), and the Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-26420007).

References

[1] Ahn, D., Choi, Y., Yu, Y., Kang, D., Choi, J.: Tuning large multimodal models for videos using reinforcement learning from ai feedback. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 923–940 (2024)

[2] An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)

[3] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

[4] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

[5] Bao, P., Shao, Z., Yang, W., Ng, B.P., Kot, A.C.: E3m: zero-shot spatio-temporal video grounding with expectation-maximization multimodal modulation. In: European Conference on Computer Vision. pp. 227–243. Springer (2024)

[6] Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Vip-llava: Making large multimodal models understand arbitrary visual prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12914–12923 (2024)

[7] Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 162–171 (2018)

[8] Chen, J., Ma, L., Chen, X., Jie, Z., Luo, J.: Localizing natural language in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8175–8182 (2019)

[9] Chen, S., Jiang, Y.G.: Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8425–8435 (2021)

[10] Chen, Z., Nie, Y., Ma, F., Xu, X., Yu, F., Long, C., et al.: Invert4tvg: a temporal video grounding framework with inversion tasks preserving action understanding ability. In: International Conference on Learning Representations. vol. 2026, pp. 125494–125513 (2026)

[11] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)

[12] Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)

[13] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16901–16911 (2024)

[14] Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence **24**(5), 603–619 (2002)

[15] Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. Advances in Neural Information Processing Systems **31** (2018)

[16] Duan, Z., Cheng, H., Xu, D., Wu, X., Zhang, X., Ye, X., Xie, Z.: Cityllava: Efficient fine-tuning for vlms in city scenario. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7180–7189 (2024)

[17] Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)

[18] Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14987–14997 (2025)

[19] Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)

[20] Gao, K., Chen, L., Niu, Y., Shao, J., Xiao, J.: Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19497–19506 (2022)

[21] Gao, M., Davis, L., Socher, R., Xiong, C.: Wslln: Weakly supervised natural language localization networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 1481–1487 (2019)

[22] Ghosh, S., Agarwal, A., Parekh, Z., Hauptmann, A.G.: Excl: Extractive clip localization using natural language descriptions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 1984–1990 (2019)

[23] Guo, C., He, Y., Nie, Y., Ma, F., Xu, X., Long, C.: T2sgrid: Temporal-to-spatial gridification for video temporal grounding. arXiv preprint arXiv:2603.06973 (2026)

[24] Guo, C., Mo, X., Nie, Y., Ma, F., Xu, X., Long, C.: Tar: Temporal anchor-constrained reasoning for video temporal grounding. In: European Conference on Computer Vision. pp. 184–201. Springer (2026)

[25] Guo, Y., Liu, J., Li, M., Liu, Q., Chen, X., Tang, X.: Trace: Temporal grounding video llm via causal event modeling. arXiv preprint arXiv:2410.05643 (2024)

[26] Hao, H., Han, J., Zhang, Y., Yue, X.: Multimodal long video modeling based on temporal dynamic context. arXiv preprint arXiv:2504.10443 (2025)

[27] He, D., Zhao, X., Huang, J., Li, F., Liu, X., Wen, S.: Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8393–8400 (2019)

[28] Hong, W., Cheng, Y., Yang, Z., Wang, W., Wang, L., Gu, X., Huang, S., Dong, Y., Tang, J.: Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8450–8460 (2025)

[29] Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)

[30] Huang, J., Li, Y., Feng, J., Wu, X., Sun, X., Ji, R.: Clover: Towards a unified video-language alignment and fusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14856–14866 (2023)

[31] Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

[32] Jin, Y., Sun, Z., Xu, K., Chen, L., Jiang, H., Huang, Q., Song, C., Liu, Y., Zhang, D., Song, Y., et al.: Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. arXiv preprint arXiv:2402.03161 (2024)

[33] Jocher, G., Qiu, J.: Ultralytics yolo11 (2024)

[34] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)

[35] Lee, J.S., Lee, S., Ahn, J., Choi, Y., Lee, J.H.: Tag: A simple yet effective temporal-aware approach for zero-shot video temporal grounding. arXiv preprint arXiv:2508.07925 (2025)

[36] Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)

[37] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

[38] Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., et al.: Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694 (2025)

[39] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

[40] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)

[41] Liao, R., Erler, M., Wang, H., Zhai, G., Zhang, G., Ma, Y., Tresp, V.: Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 6577–6602 (2024)

[42] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)

[43] Lin, Z., Zhao, Z., Zhang, Z., Wang, Q., Liu, H.: Weakly-supervised video moment retrieval via semantic completion network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11539–11546 (2020)

[44] Liu, Y., Qinghong Lin, K., Chen, C.W., Shou, M.Z.: Videomind: A chain-of-lora agent for long video reasoning. arXiv e-prints pp. arXiv–2503 (2025)

[45] Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: Debug: A dense bottom-up grounding approach for natural language video localization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5144–5153 (2019)

[46] Ma, C., Jiang, Y., Wu, J., Yuan, Z., Qi, X.: Groma: Localized visual tokenization for grounding multimodal large language models. In: European Conference on Computer Vision. pp. 417–435. Springer (2024)

[47] Meinardus, B., Rodriguez, H., Batra, A., Rohrbach, A., Rohrbach, M.: Chrono: A simple blueprint for representing time in mllms. arXiv preprint arXiv:2406.18113 (2024)

[48] Mithun, N.C., Paul, S., Roy-Chowdhury, A.K.: Weakly supervised video moment retrieval from text queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11592–11601 (2019)

[49] Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2765–2775 (2021)

[50] Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C.: X-instructblip: A framework for aligning image, 3d, audio, video to llms and its emergent cross-modal reasoning. In: European Conference on Computer Vision. pp. 177–197. Springer (2024)

[51] Pramanick, S., Mavroudi, E., Song, Y., Chellappa, R., Torresani, L., Afouras, T.: Enrich and detect: Video temporal grounding with multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24297–24308 (2025)

[52] Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435 (2024)

[53] Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. Advances in Neural Information Processing Systems **37**, 119336–119360 (2024)

[54] Qin, H., Xiao, J., Yao, A.: Question-answering dense video events. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 884–894 (2025)

[55] Qu, M., Chen, X., Liu, W., Li, A., Zhao, Y.: Chatvtg: Video temporal grounding via chat with video dialogue large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1847–1856 (2024)

[56] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

[57] Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: 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. pp. 14313–14323 (2024)

[58] Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11987–11997 (2023)

[59] Tziafas, G., Kasaei, H.: Towards open-world grasping with large vision-language models. arXiv preprint arXiv:2406.18722 (2024)

[60] Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)

[61] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

[62] Wang, Q., Yu, Y., Yuan, Y., Mao, R., Zhou, T.: Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. arXiv preprint arXiv:2505.12434 (2025)

[63] Wang, W., Huang, Y., Wang, L.: Language-driven temporal activity localization: A semantic matching reinforcement learning model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 334–343 (2019)

[64] Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)

[65] Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)

[66] Wang, Y., Meng, X., Liang, J., Wang, Y., Liu, Q., Zhao, D.: Hawkeye: Training video-text llms for grounding text in videos. arXiv preprint arXiv:2403.10228 (2024)

[67] Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)

[68] Wilkin, G.A., Huang, X.: K-means clustering algorithms: implementation and comparison. In: Second international multi-symposiums on computer and computational sciences (IMSCCS 2007). pp. 133–136. IEEE (2007)

[69] Wu, J., Li, G., Liu, S., Lin, L.: Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12386–12393 (2020)

[70] Wu, J., Zhang, Z., Xia, Y., Li, X., Xia, Z., Chang, A., Yu, T., Kim, S., Rossi, R.A., Zhang, R., et al.: Visual prompting in multimodal large language models: A survey. arXiv preprint arXiv:2409.15310 (2024)

[71] Wu, T.Y., Huang, S.Y., Wang, Y.C.F.: Data-efficient 3d visual grounding via order-aware referring. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3107–3117. IEEE (2025)

[72] Wu, Y., Wang, Y., Tang, S., Wu, W., He, T., Ouyang, W., Torr, P., Wu, J.: Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In: European Conference on Computer Vision. pp. 164–182. Springer (2024)

[73] Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., Yang, X.: Number it: Temporal grounding videos like flipping manga. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13754–13765 (2025)

[74] Xu, H., He, K., Sigal, L., Sclaroff, S., Saenko, K.: Text-to-clip video retrieval with early fusion and re-captioning. arXiv preprint arXiv:1804.05113 **2**(6), 7 (2018)

[75] Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., Du, S.: Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8978–8986 (2025)

[76] Yan, Z., Li, X., He, Y., Yue, Z., Zeng, X., Wang, Y., Qiao, Y., Wang, L., Wang, Y.: Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. arXiv preprint arXiv:2509.21100 (2025)

[77] Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)

[78] Yang, Z., Liu, Y., Hancke, G., Lau, R.W.: Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding. arXiv preprint arXiv:2509.15178 (2025)

[79] Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. AI Open **5**, 30–38 (2024)

[80] Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. Advances in Neural Information Processing Systems **32** (2019)

[81] Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9159–9166 (2019)

[82] Yue, F., Zhang, Z., Jiao, J., Liang, Z., Cao, S., Zhang, F., Shen, R.: Tempo-r0: A video-mllm for temporal video grounding through efficient temporal sensing reinforcement learning. arXiv preprint arXiv:2507.04702 (2025)

[83] Zeng, X., Li, K., Wang, C., Li, X., Jiang, T., Yan, Z., Li, S., Shi, Y., Yue, Z., Wang, Y., et al.: Timesuite: Improving mllms for long video understanding via grounded tuning. arXiv preprint arXiv:2410.19702 (2024)

[84] Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1247–1257 (2019)

[85] Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 6543–6554 (2020)

[86] Zhang, J., Guo, Y., Potamias, R.A., Deng, J., Xu, H., Ma, C.: Vtimecot: Thinking by drawing for video temporal grounding and reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24203–24213 (2025)

[87] Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12870–12877 (2020)

[88] Zhao, X., Chen, Y., Xu, S., Li, X., Wang, X., Li, Y., Huang, H.: An open and comprehensive pipeline for unified object grounding and detection. arXiv preprint arXiv:2401.02361 (2024)

[89] Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4197–4206 (2021)

[90] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974. IEEE (2024)

[91] Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)

[92] Zheng, Z., Zhang, Z., Wang, Y., Song, S., Huang, G., Yang, L.: Rethinking the architecture design for efficient generic event boundary detection. In: Proceedings of the 32nd ACM International Conference on Multimedia (2024)

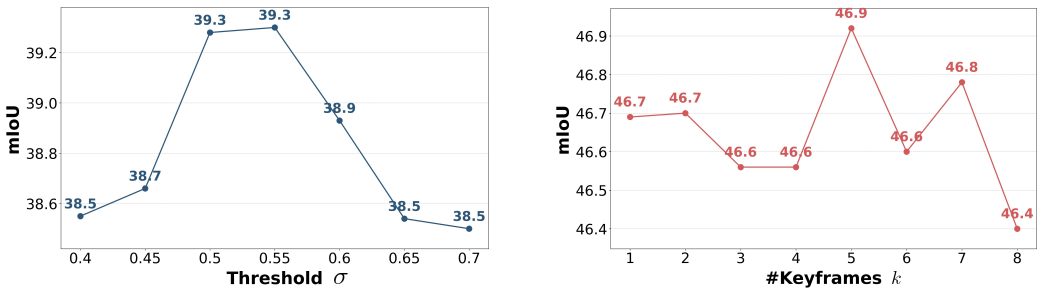
[93] Zhou, E., Qin, Y., Yin, Z., Huang, Y., Zhang, R., Sheng, L., Qiao, Y., Shao, J.: Minedreamer: Learning to follow instructions via chain-of-imagination for simulated-world control. arXiv preprint arXiv:2403.12037 (2024)

A More Implementation Details

A.1 Handling Timestamp Shift Caused by Inserting Additional Frames in ES Prompt

Qwen3-VL [3] adopts a text-timestamp alignment mechanism that matches temporal information with video frames by interleaving timestamp tags and frame tokens (e.g., <t1><frame1><t2><frame2> ...). Our ES prompt introduces additional prompt frames, which shift the original timestamp alignment (e.g., <t1><prompt frame><t2><frame1><t3><frame2> ...). To address this issue, we manually adjust the insertion of timestamp tags by removing the timestamp tag preceding each prompt frame. The resulting sequence becomes <prompt frame><t1><frame1><t2><frame2>, which preserves the correct correspondence between timestamps and video frames.

Unlike Qwen3-VL, models such as Qwen2-VL [61], LLaVA-OneVision-1.5 [2], and InternVideo2.5 [65] lack a dedicated temporal encoding mechanism and consequently cannot inherently perceive the temporal information of video frames. To equip these models with temporal localization capabilities, we follow the NumPro [73] methodology by annotating the index of each video frame in its bottom-right corner, which allows the models to utilize these frame indices as proxies for timestamps. We restrict the sequence numbering exclusively to actual video frames and exclude prompt frames from this process. As a result, the insertion of additional prompt frames does not induce a timestamp shift for these specific models.



(a) (b)

Figure 6: **Hyperparameter analysis of the proposed method.** (a) Effect of the confidence threshold σ used for filtering object detection results on mIoU when evaluating Qwen2-VL-7B on ActivityNet-Captions. The highest mIoU is achieved at a confidence threshold of 0.55. (b) Impact of the number of keyframes k selected via CLIP similarity when evaluating Qwen3-VL-8B on Charades-STA. The best performance is achieved when using 5 keyframes.

Table 7: **Additional Ablation study on different prompt combinations.**

BL	ES	OD	KM	ActivityNet-Captions R1@0.3	R1@0.5	R1@0.7	mIoU	QVHighlights R1@0.3	R1@0.5	R1@0.7	mIoU
✓				51.1	34.6	20.9	37.3	71.2	58.3	39.4	53.5
✓	✓			52.3	35.3	20.8	37.8	74.7	59.6	40.9	55.2
✓		✓		52.3	35.2	21.0	37.9	74.1	60.5	41.2	55.7
✓			✓	52.5	35.3	20.9	38.0	74.1	61.0	40.7	56.1
✓	✓	✓		52.8	35.6	21.1	38.2	75.7	60.5	41.7	56.0
✓	✓		✓	53.0	35.6	20.9	38.2	75.7	61.6	41.7	56.7
✓		✓	✓	52.9	35.5	20.9	38.2	75.5	61.9	41.7	57.0
✓	✓	✓	✓	53.2	35.7	20.9	38.3	76.3	61.9	42.2	57.0

A.2 Hyperparameter

Confidence Threshold for Object Detection. The output of an object detection model comprises a set of bounding boxes and their corresponding confidence scores. By adjusting the confidence threshold σ, we can filter not only objects with certain category labels but also morphologically similar entities, which mitigates the rigid object recognition constraints inherent in MLLMs. We evaluate the mIoU metric of our proposed OD prompt across various confidence thresholds using the Qwen2-VL-7B model [61] on the ActivityNet-Captions dataset [34]. As illustrated in Fig. 6a, the highest mIoU is achieved at a confidence threshold of 0.55.

Number of Keyframes. The proposed KM prompt identifies the top-k frames with the highest cosine similarity to the query features as keyframes and highlights them within the video sequence. We evaluate the mIoU of the KM prompt across various values of k using the Qwen3-VL-8B model [3] on the Charades-STA dataset [19]. The experimental results, as illustrated in Fig. 6b, indicate that the highest mIoU is achieved when the number of keyframes k is set to 5.

B Additional Ablation Study on Different Visual Prompt Combinations

While we provide ablation studies on various visual prompt combinations for the Charades-STA dataset in Section 4.4, we present additional results for the ActivityNet-Captions and QVHighlights datasets here. As demonstrated in Table 7, the evaluation metrics exhibit a consistent upward trend as the diversity of selected visual prompts increases. these results indicate that our selection strategy remains effective across both the ActivityNet-Captions and QVHighlights datasets.

Table 8: **Additional results on VQA benchmarks.** Our method achieves substantial accuracy improvements in tasks closely aligned with our prompt designs.

	MVBench				Video-MME			
	Action Antonym	Moving Count	State Change	Overall	Spatial Perception	Spatial Reasoning	Temporal Perception	Overall
Qwen3-VL-8B [3]	57.0	56.0	56.5	57.9	55.6	76.8	70.9	62.2
+*ours*	63.0(+6.0)	60.5(+4.5)	62.0(+5.5)	58.6(+0.7)	59.3(+3.7)	80.4(+3.6)	76.4(+5.5)	62.7(+0.5)

C Additional Results on VQA Benchmarks

To demonstrate the generalizability of our framework, we extend its application to the Video Question Answering (VQA) task and evaluate its performance on the MVBench [40] and Video-MME [17] benchmarks. During the candidate selection phase, since VQA outputs do not consist of explicit temporal intervals, it is not feasible to clip corresponding video segments for verification. Consequently, we adapt the process by inputting the various visually prompted videos along with their respective answers into the MLLM, prompting the model to verify the correctness of the answer with a binary "yes" or "no" response.

As shown in Table 8, our method achieves substantial accuracy improvements in tasks closely aligned with our prompt designs: the ES prompt enhances State Change and Temporal Perception; the OD prompt benefits Moving Count, Spatial Perception, and Spatial Reasoning; and the KM prompt significantly boosts Action Antonym. Although the overall accuracy gain is relatively modest, this stems from the fact that our three visual prompts are specifically optimized for Video Moment Retrieval (VMR) rather than the full spectrum of VQA tasks. We anticipate that redesigning visual prompts tailored to specific VQA requirements would yield even more pronounced performance gains.

D Analysis of Selection Accuracy

To analyze the accuracy of the selection mechanism, we measure the frequency of selecting the candidate with rank i, where $i \in \{1, 2, 3, 4\}$ and a higher IoU corresponds to a higher rank. As shown in the figure, the frequency of selecting the best candidate exceeds 50% on all three datasets. Among the incorrectly selected samples, candidates with higher IoU are chosen more frequently, which indicates that our method still tends to select results that are close to the ground truth even when the selection is incorrect. Since the model and the dataset annotators may have different interpretations of event boundaries, we consider such deviations to be reasonable. The combined frequency of selecting the rank 1 and the rank 2 candidates exceeds 70%. Given that the proposed selection mechanism is entirely training-free, these results demonstrate that it achieves reliable accuracy.

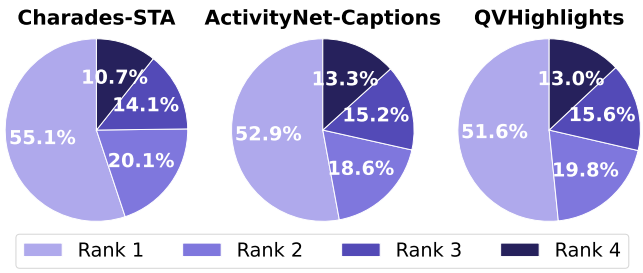


Figure 7: **Distribution of selected candidates ranked by IoU.**

E Prompt Templates

The following boxes illustrates the instruction prompt templates employed in this work.

Prompt for Stage 2: Generation of Candidate VMR Results.

Accurately pinpoint the event 'Q' in the video, answer in the format: 'From x to y'.

Prompt for Stage 3: Selection Based on Y/N Prob-Diff Score.

Does this video segment match the event 'Q'? You must only answer 'yes' or 'no'.

Prompt for extraction of salient nouns in OD prompt.

Based on the event description 'Q', your task is to capture a list of objects/targets that are within the description. These targets will be used for object detection and grounding via an object detection model. Please response in the form of a Python string list (e.g., ['object1', 'object2']).

Prompt for textual ES prompt in ablation studies.

Accurately pinpoint the event 'Q' in the video, answer in the format: 'From x to y'. Event switching at t = t_1, t_2, ... seconds.

Prompt for textual OD prompt in ablation studies.

Accurately pinpoint the event 'Q' in the video, answer in the format: 'From x to y'. I'll use the format 'frame n': [x1, y1, x2, y2] to show when and where the object appears. obj_1 appears at: 'frame 1': [x_1, y_1, x_2, y_2], 'frame 2': [x_1, y_1, x_2, y_2], ... obj_2 appears at: ...

Prompt for textual KM prompt in ablation studies.

Accurately pinpoint the event 'Q' in the video, answer in the format: 'From x to y'. The keyframes of the video are as follows: Frame n_1, Frame n_2, Frame n_3, Frame n_4, Frame n_5.

F Case Studies

As discussed in Section 1, MLLMs still encounter several limitations: (1) imprecise perception of event boundaries, (2) rigid dependency on precise object identification, and (3) inability to handle temporarily interrupted events. To address these challenges, we propose ES prompt, OD prompt and KM prompt. In this section, we provide detailed case studies to evaluate these limitations and the effectiveness of the proposed solutions.

Case Study for ES Prompt. As illustrated in Fig. 8a, for the query "Person puts on a pair of shoes", the localization result of the baseline starts from the exact moment the person's feet touch the shoes, thereby overlooking the preceding action of rising from the bed. In contrast, our proposed ES prompt explicitly partitions the video sequence into distinct events, which effectively rectifies the model perception of event boundaries.

Case Study for OD Prompt. As shown in Figure 8b, for the query "A person takes a laptop from a closet shelf", the initial baseline result deviates significantly from the ground truth, where the laptop has not even appeared at the predicted end timestamp. Upon querying the MLLM regarding the observed objects, we find that the model misidentifies the laptop as a backpack, consequently localizing only the segments related to the closet. After annotating the laptop with red bounding boxes, the model successfully recognizes the target object and yields a more accurate localization result.

Case Study for KM Prompt. As illustrated in Figure 8c, for the query "Person closing the door behind them", the baseline result only localizes the first half of the event. Observation of the video reveals that the person pauses to speak for a duration while the door is halfway closed before finally completing the action. The MLLM inaccurately perceives the closing action as concluded at the midpoint, which leads to the truncated localization. However, by highlighting the subsequent frames associated with the door closing, the model successfully localizes the entire action sequence.

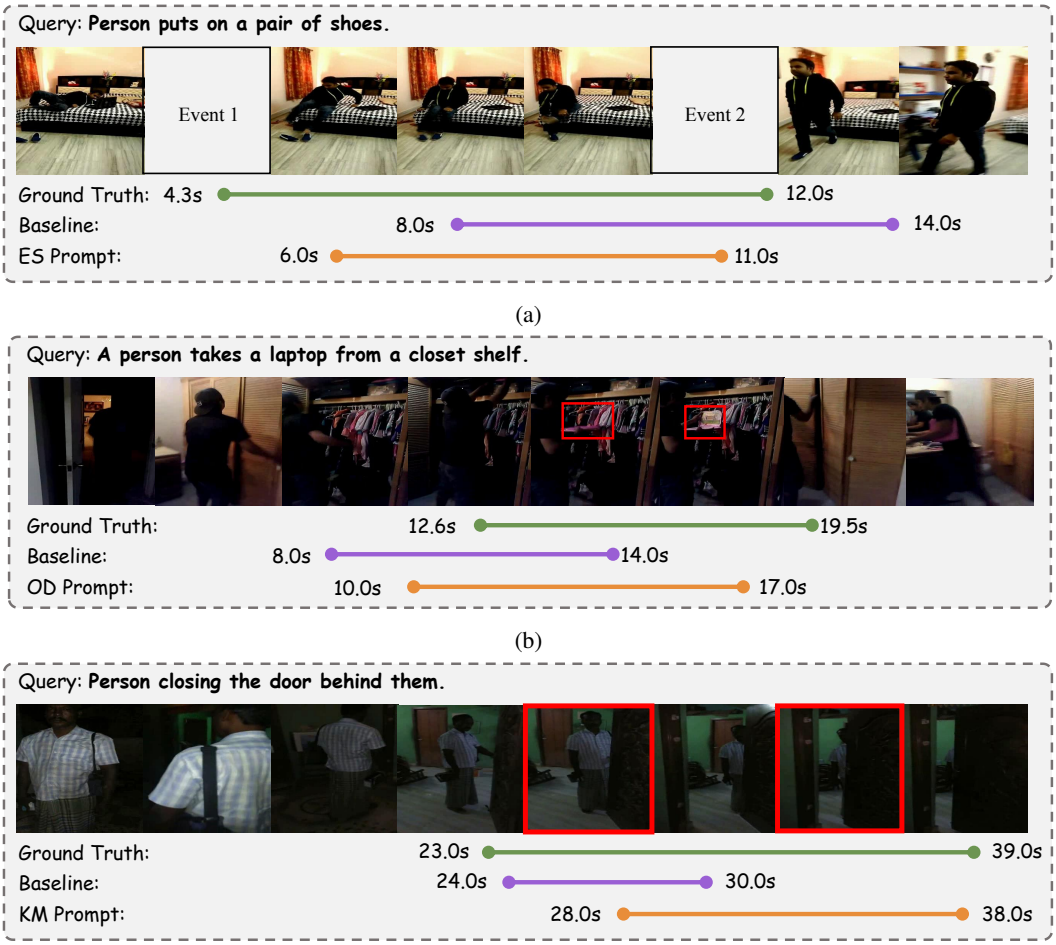


Figure 8: **Case studies for our proposed visual prompts.** (a) Baseline overlooks the preceding action of rising from the bed, while our ES prompt rectifies the model perception of event boundaries. (b) The MLLM misidentifies the laptop as a backpack, while our OD prompt directs the model's attention toward it. (c) The MLLM inaccurately perceives the closing action as concluded at the midpoint, while our KM prompt directs the model to focus on the subsequent actions.

G Broader Impacts

Positive Societal Impacts. Our work presents a training-free framework for Video Moment Retrieval (VMR), which significantly lowers the barrier to high-performance video understanding. By eliminating the need for costly fine-tuning or reinforcement learning, our method reduces the computational resources and labeled data required, making advanced video search technologies more accessible to researchers and institutions with limited budgets. This democratization of AI can facilitate applications in education (e.g., locating specific teaching moments in long lectures), accessibility (e.g., helping visually impaired users navigate video content via natural language), and digital content management, thereby promoting efficiency in handling the ever-growing volume of video data.

Negative Societal Impacts. However, like other MLLMs, our system could be misused for malicious purposes, such as generating misleading video descriptions or deepfakes if integrated into harmful pipelines. Furthermore, the model relies on pre-trained MLLMs and object detectors, which may inherit social biases present in their training data (e.g., gender or racial biases in object detection or captioning).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the main paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the extra computational cost in Section 4.6 and the suboptimal selection accuracy in Section D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: There's no theoretical results in this paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully disclose all the information needed to reproduce the main experimental results of the paper, and we will release all our code to assist the reproducibility of our experimental results. Section 4.2 and Section A contain all necessary details for reproducing our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: This paper will provides open access to the data and code, and include instructions for running the code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.2 and Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: MLLMs consume a significant amount of computational resources, making it impractical to conduct multiple experiments to obtain error bars, especially for large dataset like ActivityNet-Captions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provides computational cost information in Section 4.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We confirm the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impact in Section G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: Our study is an empirical exploration and the codes and models we release are intended solely for further research and are not meant for direct industrial application.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited and are respected for the license and terms of use explicitly mentioned.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Related codes will be released with suitable license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Yes, the paper explicitly describes the usage of MLLMs as the core component of our framework.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.