Context-Enhanced Zero-Shot Video Temporal Grounding with Adaptive Boundary Refinement

Fangkai Li[1], Hao Hu[1], Feiyu Pan[1], Yanzhen Wang[1], Yiyou Guo[2], Xiankai Lu[1,2*]
[1] *School of Software, Shandong University, Jinan, China*
[2]*School of Mathematics and Computer Science, Quanzhou Normal University, Quanzhou, China*
lfangkai@outlook.com, carrierlxk@gmail.com

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11209776

***Abstract*—In this paper, we introduce a novel training-free framework for Video Temporal Grounding (VTG) that combines pre-trained Visual Language Models (VLMs) and Large Language Models (LLMs). Existing methods often struggle with capturing the semantics of natural language queries and identifying the dynamic transitions at event boundaries. To address these challenges, our approach uses VLMs to generate detailed contextual descriptions of video content, providing richer prompts for LLMs to understand and reason about event temporal relations. Furthermore, we introduce an adaptive event boundary refinement strategy, ensuring better coverage of the full event phases. Our framework demonstrates superior performance in zero-shot settings on several benchmark datasets, including Charades-STA and ActivityNet Captions, and exhibits remarkable robustness in out-of-distribution (OOD) scenarios.**

***Index Terms*—Video temporal grounding, vision language model, prompt learning, out-of-distribution generalization**

I. Introduction

Video Temporal Grounding (VTG) [1] aims to obtain the start and end timestamps of the most semantically relevant segment in an video based on the natural language queries, serving vital roles in applications like video retrieval and video editing [2]. Most existing VTG methods [3] rely on supervised learning frameworks that require large volumes of annotated video-text pairs. However, the annotation process is labor-intensive and susceptible to human bias, limiting the scalability of these methods [4], [5]. As a result, the performance tends to degrade on unseen datasets and out-of-distribution (OOD) scenarios, limiting the generalizability [5].

Recently, visual language models (VLMs) [6]–[8] have demonstrated impressive alignment capabilities between visual and textual modalities, showing strong generalization in VTG tasks. An instinctive approach [9]–[11] is leveraging VLMs to evaluate the similarity between video proposals and queries, and then select proposals with higher similarity scores. However, VLMs are more likely to capture the climax of events in the video while neglecting the transitional phase [12], as most VLMs are pre-trained on aligned static image-text pairs or aligned video-text pairs. This is a significant challenge for VTG tasks, which require precise localization of the event's boundaries. Recent zero-shot VTG methods [4], [12] leverage large language models (LLMs) to obtain unbiased descriptions of natural language queries and event relationships, showing promising results. This indicates that the VTG task would benefit from the powerful reasoning capabilities of LLMs. However, these methods depend only on parsing natural language queries, which neglect the important cues from the video content for boundary prediction.

In response to the challenges in VTG tasks, we introduce a novel, training-free VTG framework to achieve more precise event localization. Specifically, we use the video-QA VLMs [6]–[8] to generate detailed contextual descriptions of the video content, which serve as the video context prompts for LLMs. This allows LLMs to better interpret query semantics and infer event relationships, especially for complex queries. Additionally, to enhance the model's sensitivity to temporal boundaries, we introduce an adaptive boundary refinement strategy, which generates more refined proposals around the climax proposal's endpoints, enhancing both event coverage and boundary localization accuracy.

Our method achieves performance improvements on the Charades-STA [1] and ActivityNet-Captions datasets [13]. It outperforms existing zero-shot VTG methods and delivers competitive results similar to many supervised methods while exhibiting outstanding generalization capabilities in OOD scenarios. The main contributions as three folds:

- We propose a training-free VTG framework that combines VLMs and LLMs to enhance the unbiased understanding of natural language queries and reasoning of event relationships.
- We develop an adaptive boundary refinement strategy that improves both the completeness of event coverage and the precision of boundary localization.
- Our framework outperforms previous studies on the Charades-STA and ActivityNet Captions datasets. Extensive experiments in out-of-distribution settings validate the effectiveness of our method.

II. Related Work

A. Video Temporal Grounding

Video Temporal Grounding (VTG) requires aligning textual queries with specific moments in untrimmed videos. Most

This work was supported in part by the Shandong Excellent Young Scientists Fund (ZR2024YQ006), Shandong Province Higher Education Institutions Youth Entrepreneurship and Technology Support Program (2023KJ027), the Industry-University-Research Innovation Fund for Chinese Universities-Intelligent Driving and Intelligent Cockpit Education Special Project (2024HT015) and Fujian Education and Scientific Research Project for Young and Middle-aged Teachers (JZ240047).

* Corresponding author.

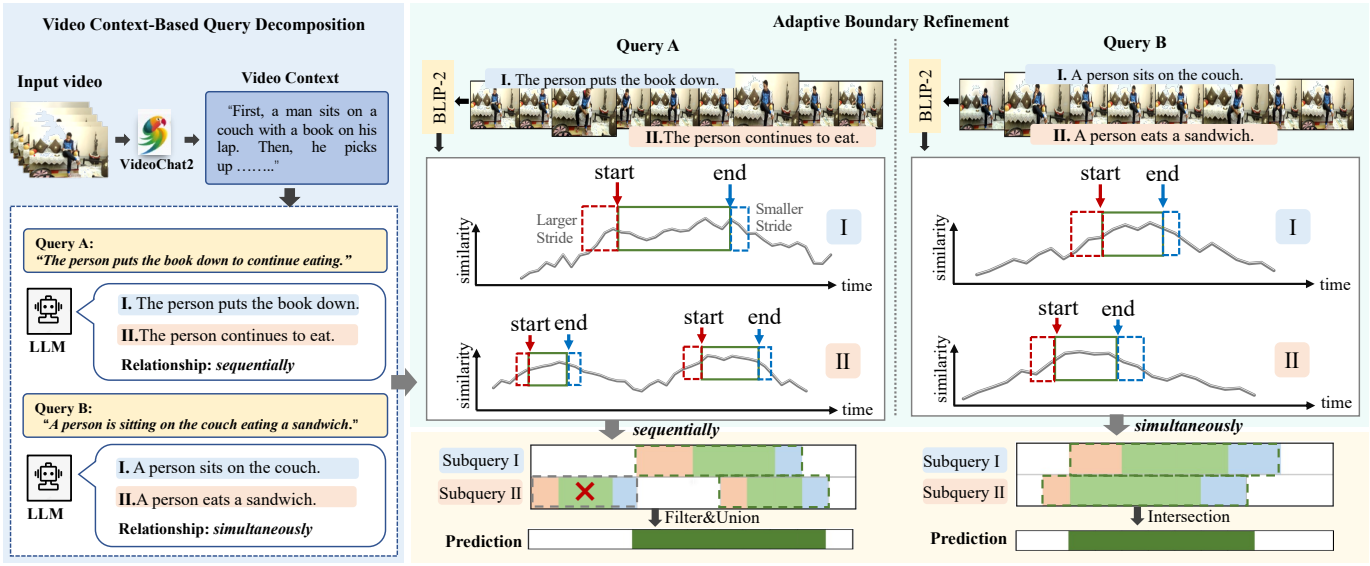

Fig. 1. Illustration of the framework. **Video Context-Based Query Decomposition** parses queries and obtains event relationships using video context; **Adaptive Boundary Refinement** enhances the event boundary location; Finally, we filter and merge proposals to obtain predictions.

fully supervised VTG methods rely on manual annotations and typically involve sophisticated multimodal interaction mechanisms [1]–[3]. Weakly supervised VTG methods relax the requirement for detailed annotations by omitting start and end timestamp annotations. For instance, CPL [14] use learnable Gaussian masks to generate hard negative proposals for weakly supervised temporal grounding. The unsupervised method like SPL [15] generates dense captions for videos and then constructs pseudo-annotation to train the model. However, the unsupervised methods still depend on specific datasets. In contrast, our method adopts strict zero-shot settings, eliminating the need for access to specific training datasets, which is crucial for enhancing the model's generalization ability.

B. VTG with Vision-Language Models

Large-scale vision-language models (VLMs) have demonstrated strong potential in multimodal alignment and generalization. Despite success in tasks like video-QA and video understanding, most VLMs still struggle with video temporal grounding, as encoding visual features into a shared language space could compromise the model's ability to recognize temporal boundaries precisely. Recent research has shifted towards training-free zero-shot VTG methods that leverage the modal alignment capability and generalization ability of VLMs, avoiding the need for training on specific datasets. For example, VTG-GPT [4] generates frame-level captions for videos using VLMs and performs temporal localization based on the similarity between captions and queries; TFVTG [12] uses the LLMs to decompose the raw query into subqueries and filter and select proposals based on order constraints. In contrast, our method provides LLMs with rich visual-semantic information generated from VLMs, enabling more unbiased parsing of complex queries and event inference, which boosts zero-shot VTG performance.

III. Method

A. Overview

Given an video V consisting of N_v frames and a natural language query q, our method aims to find the most semantically relevant segment corresponding to the query, which can be formulated as:

$$[t_s, t_e] = VTG(V, q), \tag{1}$$

where t_s and t_e denote the start and end timestamps of the time segment.

The overview of the proposed training-free framework is illustrated in Fig. 1. We begin by utilizing the video-QA VLM VideoChat2 to generate content descriptions of the input video, which serves as the video context prompts. The LLM then decomposes the query into subqueries and unbiased descriptions based on both the video context and the natural language query, while providing the event relationships. We then employ BLIP-2, an image-captioning VLM to extract the textual and frame features to calculate the frame-level similarity. Following this, we refine the scoring of the proposals based on sliding windows using an adaptive boundary refinement method. Finally, the proposals are filtered and merged according to the event relationships to obtain the final prediction.

B. Video Context-Based Query Decomposition

Natural language queries in VTG tasks often present complex structures, such as detailing various stages of an event or transitions between scenes [12]. Additionally, subjective biases in manual annotations could affect the model performance, as mentioned in [4]. Existing methods focus solely on the semantic analysis of the raw natural language query by LLMs, potentially leading to discrepancies with the video content,

such as inaccurate descriptions or improper relationships of subevents, which can introduce biases into VTG tasks.

To tackle these challenges, our method integrates the visual content comprehension of VLMs to support the LLM in performing more accurate semantic interpretation and logical reasoning for the raw queries. To elaborate, we employ VideoChat2 [6], a popular video-QA model that performs well on the MVBench. It is used to produce elaborate descriptions of the video using different prompts, including actions, event sequences, scenes, and interactions, thereby constructing the video content context:

$$c = \text{VideoChat2}(V, Prompts), \tag{2}$$

where c represents the video context provided by VideoChat2 and $Prompts$ represents random prompts of different views. The LLM then uses this contextual information to parse the raw query q in Eq 1, producing n subqueries and unbiased descriptions that better match the video content and providing the temporal relationship between the subqueries:

$$\begin{aligned} &(Q, \hat{Q}, r) = LLM(q, c), \\ &\hat{Q} = \{\hat{q}_{ij} | i = 1, ..., n; j = 1, ..., m_i\}, \end{aligned} \tag{3}$$

where n is the number of subqueries, $Q = \{q_1, ..., q_n\}$ represents the collection of subqueries, $\hat{Q}$ represents the collection of unbiased descriptions and m_i is the number of descriptions of q_i, and r represents the temporal relationship between the subqueries. As depicted in Fig. 1, for example, drawing on the video context provided by the VLM, the LLM decomposes the complex query *"The person puts the book down to continue eating"* into subqueries :

- *"The person puts the book down."*
- *"The person continues to eat."*

Meanwhile, based on the video context and the semantic logic of the query, the LLM establishes that the events occur in order of "sequentially".

C. Adaptive Boundary Refinement

After obtaining subqueries, we proceed to match these queries with the video frames. Specifically, following [11], [12], we use BLIP-2 [11] as the VLM to locate each subqueries in the video. Formally, given the video $V = \{v_1, ...v_{N_v}\}$ and a subquery's description $\hat{q}$, VLM works for projecting the features of the two modalities into unified feature space:

$$\begin{aligned} &F_V = VLM(V), F_V \in \mathbb{R}^{N_v \times d}, \\ &F_{\hat{q}} = VLM(\hat{q}), F_{\hat{q}} \in \mathbb{R}^{d}, \end{aligned} \tag{4}$$

where F_V and $F_{\hat{q}}$ denote the vision features and text features respectively, N_v is the number of video frames and d denote the feature dimension. Then we calculate their cosine similarity $S \in \mathbb{R}^{N_v}$ to measure the relevance of video frames and the description queries:

$$S = \frac{F_V \cdot F_{\hat{q}}}{\|F_V\| \|F_{\hat{q}}\|}. \tag{5}$$

As discussed in [12], the existing proposal-based methods often focus on the highest similarity regions, overlooking event boundaries. In response to this challenge, we propose an adaptive boundary refinement strategy that extends the event climax.

Climax Scoring. We first obtain proposals for the event's climax. Following the strategy proposed by SPL [15], the climax proposal for a given query should demonstrate high semantic relevance within the window and low relevance outside of it. We obtain diverse proposals using the sliding window and compute the average similarity inside and outside for each proposal. The difference between them serves as the climax score:

$$S_{climax} = \frac{1}{t_e - t_s} \sum_{t \in [t_s, t_e]} S_t - \frac{1}{N_v - (t_e - t_s)} \sum_{t \notin [t_s, t_e]} S_t. \tag{6}$$

After scoring the event climax, we apply non-maximum suppression (NMS) to discard proposals with high IoU, retaining the top-$3k$ highest-scoring proposals for following boundary refinement.

Adaptive Boundary Refinement. The boundaries of an event are typically characterized by rapid changes in cross-modal similarity. So we use the rate of change in S to quantitatively evaluate the localization of events' boundary parts. Specifically, we first apply Gaussian smoothing on S to reduce the jitter, obtaining the smoothed similarity sequence $S' = G(S)$, where G is the Gaussian filter [12]. Then, we compute the difference between consecutive values in S' as similarity difference scores:

$$D_t = \frac{|S'_t - S'_{t-1}|}{\max(|S'_t - S'_{t-1}|)}. \tag{7}$$

Next, we generate multiple candidate boundary windows around the climax proposals' endpoints for boundary refinement. Specifically, for the two endpoints t_s and t_e of a climax proposal with width w, we calculate the average similarity difference within a range of $w' = 1/2w$ centered around the t_s and t_e, which measures the average similarity difference around endpoints, take t_s as an example:

$$\text{Diff}_{t_s} = \frac{1}{w'} \sum_t D_t, t \in [t_s - 1/2w', t_s + 1/2w']. \tag{8}$$

The average difference Diff_{t_s} is then utilized to guide the adaptive adjustment of the boundary window size and stride:

$$\begin{aligned} &\text{size} = \text{round}(\text{size}_0 \times (1 - \gamma \cdot \text{Diff}_{\text{t}_\text{s}}), \\ &\text{stride} = \text{round}(\text{stride}_0 \times (1 - \gamma \cdot \text{Diff}_{\text{t}_\text{s}}), \end{aligned} \tag{9}$$

where $\text{size}_0 = \text{round}(\text{w}'/2)$ and $\text{stride}_0 = \text{round}(\text{w}'/4)$ are the default window size and stride, γ is the hyperparameter, and rand is the rounding function. Boundary proposals are generated using the adaptive window size and stride, enabling smaller windows in regions with rapid similarity changes for more precise proposals. For boundary proposals' scoring, we consider both the similarity score S_t and the similarity difference D_t at each timestamp to get boundary scores $\hat{S}$,

TABLE I
IID TESTING RESULTS ON CHARADES-STA [1] AND ACTIVITYNET-CAPTIONS [13]. SEE §IV-B FOR DETAILS.

Method	Reference	Setting	Charades-STA				ActivityNet-Captions			
			R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN [3]	AAAI'20	Fully -Supervised	57.31	45.75	27.88	41.05	60.32	43.41	25.04	42.45
MMN [2]	AAAI'22		65.43	53.25	31.42	46.46	**64.48**	**48.24**	**29.35**	**46.61**
EMB [16]	ECCV'22		**72.50**	**58.33**	**39.25**	**53.09**	64.13	44.81	26.07	45.59
VCA [17]	MM'21	Weakly -Supervised	58.58	38.13	19.57	38.49	50.45	31.00	-	33.15
CPL [14]	CVPR'22		66.09	50.06	22.83	43.56	55.73	31.37	13.68	36.65
Huang et al. [18]	CVPR'23		**69.16**	**52.18**	**23.94**	**45.20**	**58.07**	**36.91**	-	**41.02**
PSVL [19]	ICCV'21	Unsupervised	46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
PZVMR [20]	MM'22		46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al. [21]	WACV'23		52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
SPL [15]	ACL'23		**58.92**	**41.16**	**21.21**	**40.41**	**50.24**	27.24	15.03	**35.44**
VideoChat-7B [6]	CVPR'24	Zero-Shot	9.0	3.3	1.3	6.5	8.8	3.7	1.5	7.2
VideoChatGPt-7B [8]	ACL'24		20.1	7.7	1.7	13.7	26.4	13.6	6.1	18.9
VTG-GPT† [4]	APP.SCI.'24		59.48	43.68	**25.94**	39.81	47.13	28.25	12.84	30.49
TFVTG† [12]	ECCV'24		67.04	49.97	24.32	44.51	**49.34**	27.02	13.39	34.10
Ours†	——		**67.91**	**50.15**	24.89	**45.41**	49.23	**27.54**	**13.81**	**34.47**

"†" denotes the **Training-Free** methods under the zero-shot setting.

TABLE II
RESULTS ON CHARADES-STA AND ACTIVITYNET-CAPTIONS UNDER TEMPORAL ANNOTATION SHIFT. SEE §IV-C FOR DETAILS.

Method	Setting	Charades-STA OOD(10 seconds)			ActivityNet-Captions OOD(30 seconds)		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
2D-TAN [3]	Fully	27.1	13.1	25.7	16.4	6.6	23.2
VDI [22]		25.9	11.9	26.7	**20.9**	7.1	**27.6**
DCM [23]		**44.4**	**19.7**	**42.3**	18.2	**7.9**	24.4
CNM [24]	Weakly	9.9	1.7	21.6	**6.1**	0.4	21.0
CPL [14]		**29.9**	**8.5**	**32.2**	4.7	**0.5**	**21.1**
Luo et al. [10]	Zero-Shot	40.3	18.2	38.2	18.4	6.8	21.1
TFVTG [12]		45.9	20.8	43.0	20.4	11.2	31.7
Ours		**47.0**	**22.3**	**44.4**	**21.4**	**11.5**	**32.4**

and calculate the difference of the average boundary score inside and outside the window:

$$\hat{S}_t = \alpha \cdot S_t + (1-\alpha) \cdot D_t,$$
$$S_{boundary} = \frac{1}{t_e - t_s} \sum_{t \in [t_s, t_e]} \hat{S}_t - \frac{1}{N_v - (t_e - t_s)} \sum_{t \notin [t_s, t_e]} \hat{S}_t, \tag{10}$$

where $S_{boundary}$ is the boundary scores, and α is the hyperparameter used to balance the impact of two types of scores. At endpoints of the $3k$ climax proposals, we combine the highest-scoring boundary window with the climax window to generate a complete proposal. Finally, the top-k proposals are selected based on the sum of climax scores and boundary scores:

$$S_{total} = S_{climax} + S_{boundary}. \tag{11}$$

D. Proposal Filtering and Merging

Following [12], we filter and integrate the proposals of subqueries based on the temporal relationship provided by the LLM. For a raw query, we enumerate all possible combinations of its subqueries' proposals, and all invalid combinations that fail to satisfy the temporal order constraints are discarded. For the remaining combinations, we calculate the total score of proposals, and the highest score combination will be selected. Finally, proposals in the combination are merged based on their relationships: intersections for "simultaneously" and unions for "sequentially". Notably, our proposed method leverages the cross-modal alignment capability of pre-trained large models to perform VTG tasks, without requiring additional training.

IV. EXPERIMENTS

We conducted extensive comparisons with SOTA methods including supervised, weakly supervised, unsupervised, and zero-shot methods.

A. Experimental Settings

Datasets. To verify the effectiveness of the proposed framework, we conducted extensive experiments on two datasets.

- **Charades-STA [1]:** This dataset is constructed from the Charades dataset. It includes 16,128 annotations describing indoor activities, with 3,720 annotations for testing.
- **ActivityNet Captions [13]:** Derived from the ActivityNet dataset, this dataset consists of 2,000 videos spanning multiple domains and 71,957 queries.

Evaluation Metrics. Following the previous work [14], [19], [24], we adopted Recall-1 at Intersection over Union (IoU) thresholds m (R1@m) and mean Intersection over Union (mIoU) as the evaluation metrics. Specifically, R1@m measures the percentage of top-ranked predicted segments with IoU exceeding the threshold m, while mIoU represents the average IoU across all test samples.

Implementation Details. We take BLIP-2 [11] as the vision-language model to extract video frame features, downsampling the input videos to 3 FPS. GPT-4 Turbo serves as the LLM to generate fine-grained subqueries and event

TABLE III
EVALUATIONS RESULTS OF COMPOSITIONAL GENERALIZABILITY ON CHARADES-STA [1]. SEE §IV-D FOR DETAILS.

Method	Setting	Charades-CG					
		Novel-Composition			Novel-Word		
		R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
2D-TAN [3]	Fully	30.91	12.23	29.75	29.36	13.21	28.47
SCDM [25]		27.73	12.25	30.84	-	-	-
CPL [14]	Weakly	39.11	15.60	35.53	45.90	22.88	-
Luo et al. [10]	Zero-Shot	40.27	16.27	-	45.04	21.44	-
TFVTG [12]		43.84	18.68	40.19	56.26	28.49	46.90
Ours		**45.28**	**18.80**	**41.20**	**57.23**	**28.71**	**47.29**

TABLE IV
RESULTS ON CHARADES-CD DATASET [26]. REFER TO §IV-C.

Method	Setting	R1@0.3	R1@0.5	R1@0.7
2D-TAN [3]	Fully	43.45	30.77	11.75
SCDM [25]		52.38	41.60	22.22
SPL [15]	Unsupervised	62.96	38.25	15.53
TFVTG [12]	Zero-Shot	65.07	49.24	23.05
Ours		**67.04**	**49.62**	**23.53**

TABLE V
ABLATIONS ON EACH COMPONENT. REFER TO §IV-F.

	LLM	Video Context	Boundary Refine	mIoU
①	✗	✗	✗	38.83
②	✗	✗	✓	39.48
③	✓	✗	✗	42.77
④	✓	✗	✓	43.22
⑤	✓	✓	✗	44.04
⑥	✓	✓	✓	**45.41**

TABLE VI
ABLATIONS ON VIDEO-QA VLMS. REFER TO §IV-F.

Video-QA Models	R1@0.5	R1@0.7	mIoU
VideoLLaMA [7]	43.70	19.32	42.89
InternVideo [9]	45.62	20.54	43.91
VideoChat2 [6]	**50.15**	**24.89**	**45.41**

relationships. For the hyperparameter, we set $\gamma = 0.5$ in Eq. 9, $\alpha = 0.7$ in Eq. 10 and $k = 3$.

B. Performance under IID Settings

Under the independent identically distribution (IID) setting, we evaluated the proposed method on the official splits of the Charades-STA [1] and ActivityNet Captions [13] datasets and compared it with existing VTG methods, including supervised, weakly supervised, unsupervised, and zero-shot approaches. As shown in Table I, our method demonstrates outstanding performance in the zero-shot setting, with the mIoU of 45.41% on Charades-STA and 34.47% on ActivityNet Captions. Notably, it outperforms most weakly supervised and unsupervised methods and even surpasses some fully supervised methods. These highlight the effectiveness of our method in leveraging the cross-modal alignment capabilities of VLMs.

C. OOD Evaluations with Distribution Shifts

To demonstrate the generalization ability of our method, particularly its robustness under distributional shifts, we perform two types of out-of-distribution (OOD) experiments:

- **Temporal Annotation Shift:** Following the [23], we add randomly generated irrelevant video segments (10 seconds for Charades-STA and 30 seconds for ActivityNet Captions) to the beginning of each video, and modify the annotations accordingly. As reported in Table II, for Charades-STA, our method surpasses the latest zero-shot method [12] with +1.4% on mIoU. For ActivityNet Captions, following DCM [23] to remove long moments for fair comparison, we obtain the mIoU of 32.4%, setting a new SOTA in zero-shot methods.
- **Charades-CD [26] Dataset Evaluation:** This dataset is created by re-partitioning the training and testing sets of the original dataset to change the distribution. As reported in Table IV, most fully supervised and unsupervised methods witness a significant performance decline. In contrast, our method surpasses the fully supervised method [25] by 14.6% and unsupervised method [15] by 4.1% on R1@0.3, and it also outperforms the latest zero-shot supervised method [12] by 1.97% on R1@0.3, proving its superior generalization ability.

D. Compositional Generalizability Evaluations

To further investigate the model's adaptability to queries from different sources, we evaluate its compositional generalizability on the Charades-CG dataset with two experimental settings proposed in [5]:

- **Novel-Composition:** Queries involve novel combinations of vocabulary.
- **Novel-Word:** Queries contain entirely new words.

As shown in Table III, our method achieves the best performance with the mIoU of 41.20% under the Novel-Composition setting and 47.29% under the Novel-Word setting, showing the advanced compositional generalizability.

E. Inference Time Comparison

Our method is training-free, requiring only inference. As shown in Tableb VII, the VLM and LLM introduce a slight inference time increase, but it critically enhances the context understanding, thereby boosting performance (mIoU +5.9 on Charades-STA).

TABLE VII
INFERENCE TIME ON CHARADES-STA DATASET [1]

Method	mIoU	Avg infer time (second)
w/o LLM&VLM	39.48	11.3
w/o VLM	43.22	12.1
Full model	**45.41**	13.7

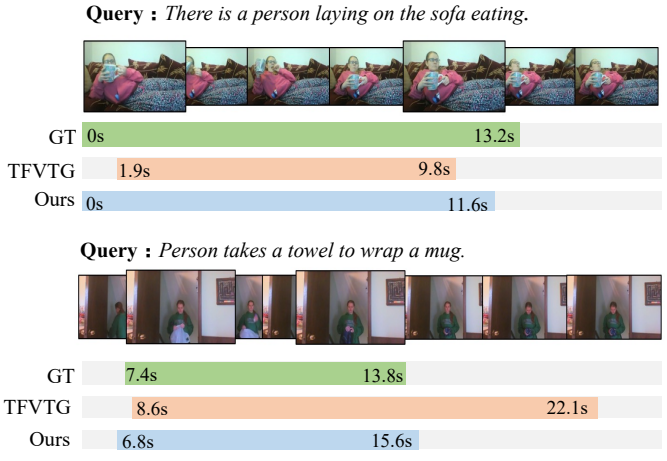

Fig. 2. Qualitative results on Charades-STA [1].

F. Ablation Study

To demonstrate the effectiveness of each module in our framework, we conduct comprehensive ablation studies.

Impact of each component. Table V presents the effectiveness of each component in our framework. With *LLM* disabled, we directly extract textual features of the raw query. With *Video Context* disabled, the LLM generates subqueries and descriptions solely from the raw query. With *Adaptive Boundary Refinement* disabled, we use the climax proposals directly as the candidate proposals. It is clear that all components significantly contribute to the model's performance.

Ablations on Video-QA VLMs. Table VI presents the performance of different video-QA VLMs in §III-B, including VideoLLaMA [7], InternVideo [9], and VideoChat2 [6]. VideoChat2 achieves the best performance.

G. Qualitative Results

Fig. 2 illustrates the qualitative results on the Charades-STA dataset. Our method provides more precise coverage of entire events and exhibits greater sensitivity to event boundaries.

Conclusion

In this paper, we introduced an innovative training-free zero-shot VTG framework that integrates the capabilities of VLMs and LLMs. By generating video context descriptions, we enhance the LLM's reasoning ability of the natural language query, while the adaptive boundary refinement strategy improves the precision of event boundary localization. Experimental results demonstrate that our method surpasses most existing methods on benchmark datasets while exhibiting strong robustness and generalization in OOD settings.

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia, "Tall: Temporal activity localization via language query," in *ICCV*, 2017, pp. 5267–5275.

[2] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu, "Negative sample matters: A renaissance of metric learning for temporal grounding," in *AAAI*, 2022, vol. 36, pp. 2613–2623.

[3] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *AAAI*, 2020, vol. 34, pp. 12870–12877.

[4] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du, "Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt," *Applied Sciences*, vol. 14, no. 5, pp. 1894, 2024.

[5] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang, "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *CVPR*, 2022, pp. 3032–3041.

[6] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al., "Mvbench: A comprehensive multi-modal video understanding benchmark," in *CVPR*, 2024, pp. 22195–22206.

[7] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint:2311.10122*, 2023.

[8] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *ACL*, 2024, pp. 12585–12602.

[9] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al., "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint:2307.06942*, 2023.

[10] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu, "Zero-shot video moment retrieval from frozen vision-language models," in *WACV*, 2024, pp. 5464–5473.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*. PMLR, 2023, pp. 19730–19742.

[12] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu, "Training-free video temporal grounding using large-scale pre-trained models," in *ECCV*, 2024, pp. 20–37.

[13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.

[14] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu, "Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning," in *CVPR*, 2022, pp. 15555–15564.

[15] Minghang Zheng, Shaogang Gong, Hailin Jin, Yuxin Peng, and Yang Liu, "Generating structured pseudo labels for noise-resistant zero-shot video sentence localization," in *ACL*, 2023, pp. 14197–14209.

[16] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu, "Video activity localisation with uncertainties in temporal boundary," in *ECCV*. Springer, 2022, pp. 724–740.

[17] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang, "Visual co-occurrence alignment learning for weakly-supervised video moment retrieval," in *ACM MM*, 2021, pp. 1459–1468.

[18] Yifei Huang, Lijin Yang, and Yoichi Sato, "Weakly supervised temporal sentence grounding with uncertainty-guided self-training," in *CVPR*, 2023, pp. 18908–18918.

[19] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi, "Zero-shot natural language video localization," in *ICCV*, 2021, pp. 1470–1479.

[20] Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan, "Prompt-based zero-shot video moment retrieval," in *ACM MM*, 2022, pp. 413–421.

[21] Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn, "Language-free training for zero-shot video grounding," in *WACV*, 2023, pp. 2539–2548.

[22] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu, "Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training," in *CVPR*, 2023, pp. 23045–23055.

[23] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua, "Deconfounded video moment retrieval with causal intervention," in *ACM SIGIR*, 2021, pp. 1–10.

[24] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu, "Weakly supervised video moment localization with contrastive negative sample mining," in *AAAI*, 2022, vol. 36, pp. 3517–3525.

[25] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *NeurIPS*, 2019, vol. 32.

[26] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu, "A closer look at temporal sentence grounding in videos: Dataset and metric," in *Proceedings of the 2nd international workshop on human-centric multimedia analysis*, 2021, pp. 13–21.