Zero-shot Video Moment Retrieval via Off-the-shelf Multimodal Large Language Models

Yifang Xu[1], Yunzhuo Sun[2], Benxiang Zhai[1], Ming Li[1], Wenxin Liang[2], Yang Li[1], Sidan Du[1]

[1]Nanjing University
[2]Dalian University of Technology
{xyf, zbx, liming}@smail.nju.edu.cn, {sunyunzhuo, wxliang}@mail.dlut.edu.cn, {yogo, coff128}@nju.edu.cn

Abstract

The target of video moment retrieval (VMR) is predicting temporal spans within a video that semantically match a given linguistic query. Existing VMR methods based on multimodal large language models (MLLMs) overly rely on expensive high-quality datasets and time-consuming fine-tuning. Although some recent studies introduce a zero-shot setting to avoid fine-tuning, they overlook inherent language bias in the query, leading to erroneous localization. To tackle the aforementioned challenges, this paper proposes **Moment-GPT**, a tuning-free pipeline for zero-shot VMR utilizing frozen MLLMs. Specifically, we first employ LLaMA-3 to correct and rephrase the query to mitigate language bias. Subsequently, we design a span generator combined with MiniGPT-v2 to produce candidate spans adaptively. Finally, to leverage the video comprehension capabilities of MLLMs, we apply Video-ChatGPT and span scorer to select the most appropriate spans. Our proposed method substantially outperforms the state-of-the-art MLLM-based and zero-shot models on several public datasets, including QVHighlights, ActivityNet-Captions, and Charades-STA.

1 Introduction

Video moment retrieval (VMR) is a crucial task in the field of video understanding, attracting widespread attention in the last few years owing to its potential applications in video surveillance (Lyu and Zhang 2023) , human-computer interaction (Yan et al. 2024), etc. It aims to locate temporal spans (segments) that are semantically related to a specified sentence query from an untrimmed video, with each span comprising a beginning and an ending moment. Fig. 1 (a) presents an instance of VMR.

Recently, large language models (LLMs), like GPT-4 (OpenAI 2023) and LLaMA-3 (AI@Meta 2024), have attained noteworthy success in the natural language processing (NLP) domain. This advancement facilitates the development of multimodal LLMs (MLLMs) (Chen et al. 2023; Maaz et al. 2023) in visual and multimodal domains. Most recent studies (Huang et al. 2023a; Ren et al. 2023) demonstrate that training only LoRA (Adapter) (Hu et al. 2021) can empower MLLMs to seize spans, as depicted in Fig. 1 (b). However, these MLLM-based methods necessitate intricate multi-stage fine-tuning strategies tailored for VMR. Moreover, they depend on large video datasets annotated with high-quality spans and queries, which are time-consuming and costly to collect.

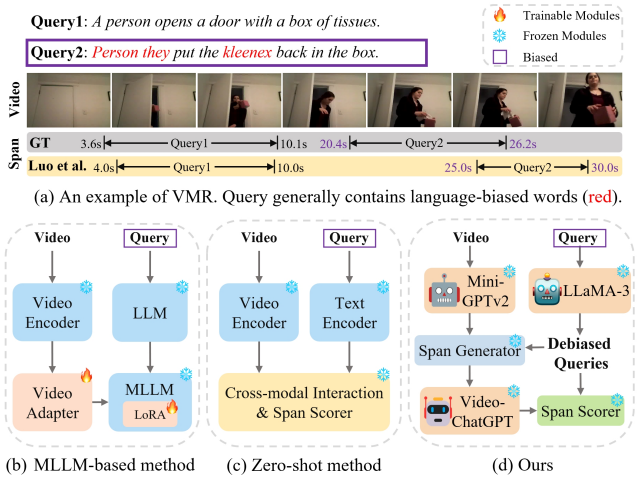


Figure 1: (a) The query containing language bias results in erroneous localization. (b) MLLM-based method demand fine-tuning using large amounts of data. (c) Zero-shot method cannot avoid performance degradation caused by language bias. (d) Our proposed Moment-GPT harnesses MLLMs without further fine-tuning. It also utilizes LLM to reduce bias, enhancing overall accuracy.

To alleviate the burden of manual annotations, some previous works (Luo et al. 2023; Wattasseril et al. 2023) propose a zero-shot setting for VMR, which simply reuses off-the-shelf MLLMs trained on other tasks, as shown in Fig. 1 (c). Nevertheless, these zero-shot methods neglect language biases (Linell 2004; Liang et al. 2020) in human-annotated queries, including (1) rare words (2) spelling and grammatical errors. As exemplified in `query2` of Fig. 1 (a), "*kleenex*" is a rare word compared to "*tissues*", which causes the model (Luo et al. 2023) to favor common words ("*person*", "*put*", "*box*"), resulting in inaccurate localization. Additionally, grammatical errors ("*Person they*") can misguide the model because only one person appears in the video. Therefore, it is necessary to eliminate the above biases to improve accuracy.

To tackle the above challenges, this paper proposes a zero-shot VMR framework, **Moment-GPT**, employs frozen

MLLMs and without additional fine-tuning on VMR data, as illustrated in Fig. 1 (d). To mitigate language biases, we utilize LLaMA-3 (AI@Meta 2024) to optimize the raw query, yielding debiased queries. Given that video contains more redundant information than highly generalized text (Wu et al. 2022), and inspired by humans linguistically understanding videos (Barrett et al. 2015; Rohrbach et al. 2013), we apply MiniGPT-v2 (Chen et al. 2023) to obtain frame-level captions from the input video. Then, we compute the similarities between these captions and debiased queries, adaptively producing candidate spans through the proposed span generator. To leverage the video understanding capabilities of MLLMs, and considering that the existing MLLMs (Maaz et al. 2023; Zhang et al. 2023) are better at video captioning than VMR, we use Video-ChatGPT (Maaz et al. 2023) to generate span-level captions. Finally, the span scorer calculates the relevance between span-level captions and debiased queries, followed by post-processing to obtain the final results. To summarize, our main contributions include:

- We propose **Moment-GPT**, a zero-shot VMR approach using off-the-shelf MLLMs for direct inference.
- We devise a new strategy for query debiasing utilizing LLaMA-3 to enhance performance. In addition, we design the span generator and span scorer to exploit the visual comprehension abilities of MiniGPT-v2 and Video-ChatGPT effectively.
- Extensive experimental results demonstrate that our method outperforms the SOTA MLLM-based and zero-shot approaches on three VMR datasets. Significantly, it also exceeds most supervised models.

2 Related Work

Video moment retrieval. VMR constitutes a promising yet challenging task emphasizing retrieving relevant spans from a video, given a linguistic query. Existing fully-supervised VMR approaches (Sun et al. 2023; Xu et al. 2024a,b; Lei et al. 2021) conventionally hinge on extensive datasets annotated with queries and corresponding spans for training. However, manually collecting VMR data is costly and labor-intensive; for example, producing QVHighlights (Lei et al. 2021) took around $17,000 and 1,500 hours. To alleviate reliance on spans, prior studies (Zheng et al. 2022a,b) propose a weakly-supervised setup to learn the unmatched video-query pairs. Further diminishing the dependency on queries, some works (Nam et al. 2021; Wang et al. 2022a) introduce unsupervised framework, leveraging k-means clustering or CLIP (Radford et al. 2021) to generate pseudo queries from videos, or select from a query database. Please note that we align with recent works (Diwan et al. 2023; Luo et al. 2023) and categorize partially zero-shot methods (Nam et al. 2021; Wang et al. 2022a) as unsupervised.

Zero-shot video moment retrieval. To reduce the burden of manual annotation and circumvent vision bias from specific VMR videos, recent works (Diwan et al. 2023; Luo et al. 2023) propose a zero-shot setting repurposing frozen models pretrained on other tasks and without any fine-tuning. Luo et al. (Luo et al. 2023) generate spans using InternVideo (Wang et al. 2022b) with refined masks and clustering. Diwan et al. (Diwan et al. 2023) and Wattasseril et al. (Wattasseril et al. 2023) apply a shot-detection technique and CLIP (BLIP-2 (Li et al. 2023a)) for span computation, but this strategy is not suitable for scenarios with rapid shot transitions, thereby causing poor overall localization effect. In addition, the above zero-shot methods overlook language bias in the original queries, leading to highly inaccurate predictions on biased queries, as depicted in Fig. 1 (a). To solve this problem, this paper designs a debiasing strategy using LLM (AI@Meta 2024) to reduce the bias.

Multimodal large language models. Recent LLMs (Xu et al. 2023; OpenAI 2023; Touvron et al. 2023; AI@Meta 2024) have attracted widespread attention from researchers due to their remarkable success in various NLP tasks. This success promotes the development of MLLMs (Zhu et al. 2023; Chen et al. 2023; Liu et al. 2023) in the field of computer vision, with representative works like LLaVA (Liu et al. 2023) and MiniGPT-v2 (Chen et al. 2023) leveraging appropriate prompts to engage in dialogue, enabling them to summarize images and generate detailed textual descriptions. Subsequently, researchers expand single-frame images into multi-frame videos, with methods like VideoLLMA (Zhang et al. 2023) and Video-ChatGPT (Maaz et al. 2023) demonstrating robust video understanding capabilities, yielding superior zero-shot results in tasks including video captioning and video Q&A. However, these MLLMs (Zhang et al. 2023; Maaz et al. 2023) exhibit poor performance in VMR due to the absence of span constraints during training. To address this, recent works (Ma et al. 2023; Li et al. 2024) devise complex multi-stage strategies to train LoRA, endowing MLLMs with temporal localization abilities. Nonetheless, these training strategies are time-consuming and require laborious collection of precise annotations. Moreover, to enhance localization accuracy, some works (Huang et al. 2023a; Ren et al. 2023) employ a sliding window to pre-generate overlapping candidate spans, significantly increasing computational costs. Unlike the above methods, this paper leverages frozen MLLMs (AI@Meta 2024; Chen et al. 2023; Maaz et al. 2023) to address the above challenges.

3 Method

3.1 Overview

Given an untrimmed video $V = \{v_i\}_{i=1}^{L_v}$ containing L_v frames, and a textual query $Q = \{q_i\}_{i=1}^{L_q}$ comprising L_q words, the goal of video moment retrieval (VMR) is to predict all temporal spans (segments) $T = \{t^s, t^e\} \in \mathbb{R}^{N_t \times 2}$ semantically relevant to the query: $T = \text{VMR}(V, Q)$, in which t^s and t^e represent the start and end moments, respectively.

Fig. 2 illustrates the architecture of our proposed Moment-GPT. Concretely, we first apply LLaMA-3 (AI@Meta 2024) to inspect and rephrase Q, thereby mitigating language bias and generating debiased queries $D \in \mathbb{R}^{N_d \times L_d}$. Subsequently, MiniGPT-v2 (Chen et al. 2023) summarizes each image to produce frame-level captions $C^f \in \mathbb{R}^{L_v \times L_f}$. The frame scorer computes frame-wise similarities $S^f \in \mathbb{R}^{N_d \times L_v}$ between D and C^f, and then span generator dynamically constructs

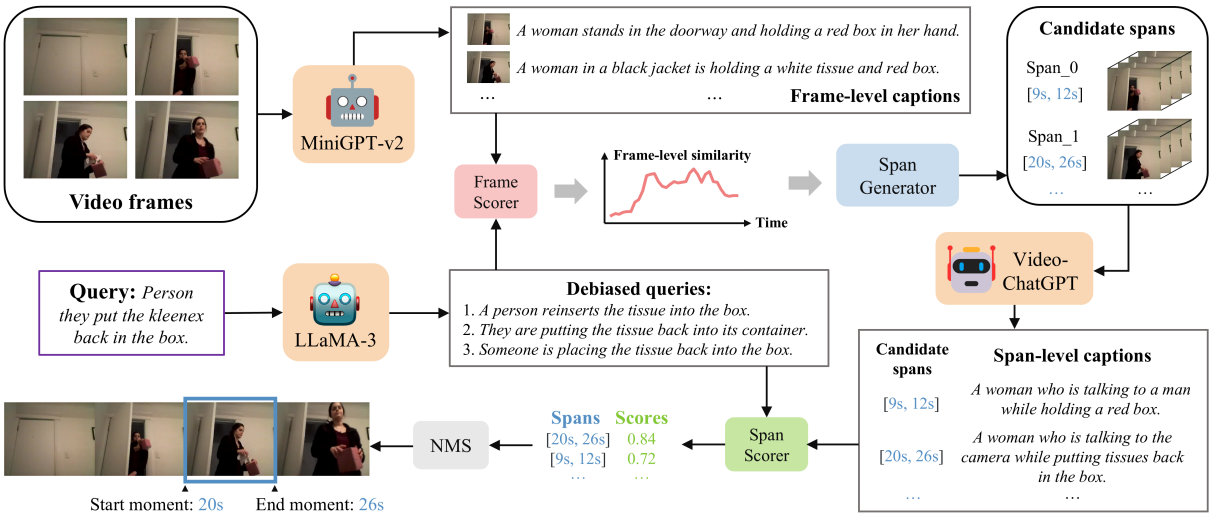


Figure 2: The overall architecture of Moment-GPT. It first utilizes LLaMA-3 to reduce language bias (Sec. 3.2). Next, construct candidate spans by MiniGPT-v2, frame scorer, and span generator (Sec. 3.3). Finally, select the most relevant spans using Video-ChatGPT, span scorer, and NMS (Sec. 3.4).

candidate spans $T^p \in \mathbb{R}^{N_p \times 2}$ from S^f. To leverage video understanding abilities of MLLMs, we feed candidates T^p into Video-ChatGPT (Maaz et al. 2023) to obtain span-level captions $C^s \in \mathbb{R}^{N_p \times L_s}$. After that, the span scorer calculates span-level score $S \in \mathbb{R}^{N_p}$ based on relevance between C^s and D, followed by non-maximum suppression (NMS) to derive the most accurate spans $T \in \mathbb{R}^{N_t \times 2}$.

3.2 Reducing Language Bias

Human-annotated queries often contain language bias stemming from the annotators' subjectivity, such as rare words, spelling and grammatical mistakes, which result in the model locating erroneous spans. To address these, we adopt a human-like approach (Winograd 1972; Liddy 2001) to natural language processing, first using LLaMA-3 (AI@Meta 2024) to rectify spelling and grammar mistakes in the raw query. Then, synonym substitution (Mekala et al. 2023; Kong et al. 2024) is employed to rewrite the rectified query, reducing the occurrence of rare words. However, we observe that very few queries still have rare words after rewriting. Thus, we task LLaMA-3 with emulating diverse expression styles to generate multiple rewritten queries with unchanged semantics, minimizing the likelihood of encountering rare words. Finally, we consolidate the above ideas and craft the following prompt: `"Raw sentence: '<Query>' \n \n Task 1: Please detect and rectify spelling and grammatical mistakes in the raw sentence. \n Task 2: Please rewrite the rectified sentence using different wording while ensuring that the rewritten sentence retains the original meaning. Please provide three different rewrites. Please avoid rare words and phrases. \n \n Please only return the rewritten sentences."` Here, `<Query>` represents the raw query. We refer to the text returned by LLaMA-3 as debiased queries $D \in \mathbb{R}^{N_d \times L_d}$. Fig. 3 depicts the process of query debiasing, where grammatical mistakes ("*Person they*") are rectified to more accurate terms such as "*A person*", "*They*", or "*Someone*". Similarly, the rare word ("*kleenex*") is substituted with a more common alternative ("*tissues*"). Fig. 6 (top) demonstrates that utilizing these debiased queries can get more precise results.

3.3 Generating Candidate Spans

To improve localization accuracy, existing methods (Huang et al. 2023a; Ren et al. 2023) apply the sliding-window strategy to construct candidate spans. However, this strategy often leads to the creation of excessively overlapping candidates, resulting in significant computational overhead. In response to this challenge, we propose the following solution (image captioning, frame scorer, and span generator) to adaptively generate candidates $T^p \in \mathbb{R}^{N_p \times 2}$ with low overlap.

Image captioning. Drawing from previous zero-shot method (Diwan et al. 2023; Wattasseril et al. 2023), in our ini-

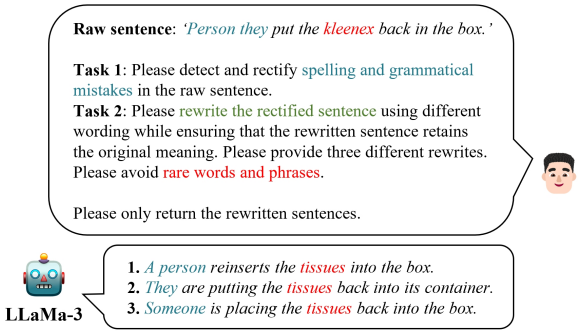
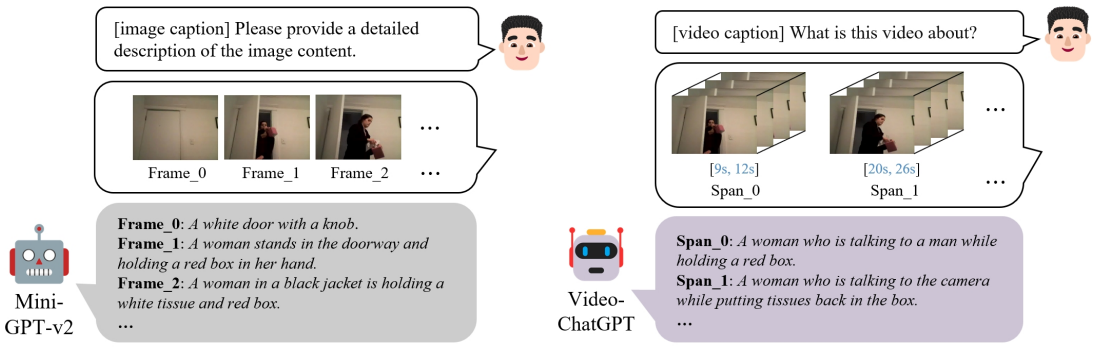


Figure 3: Reduce language bias in raw query via LLaMA-3. Bold, italics, and colored fonts are utilized only for presentation and are not employed in the code.



(a) Image captioning (b) Video captioning

Figure 4: (a) Image captioning via MiniGPT-v2. (b) Video captioning via Video-ChatGPT. Frame_N and Span_N are just for demonstration convenience and do not exist in reality.

tial experiments, we use BLIP-2 (Li et al. 2023a) to calculate S^f between D and each frame in the video. However, this approach is susceptible to visual redundancy (Wu et al. 2022; Yu et al. 2024), resulting in suboptimal outcomes (see Tab. 5, row 1). Inspired by human comprehension of visual signals through language (Barrett et al. 2015; Rohrbach et al. 2013), we utilize MiniGPT-v2 to transform images into highly informative captions $C^f \in \mathbb{R}^{L_v \times L_f}$, as shown in Fig. 4 (a). The instruction as follows: "`[image caption] Please provide a detailed description of the image content. <Image>`". Here, `[]` is used to emphasize the task within the square brackets, and `<Image>` denotes the actual image.

Frame scorer. Considering that BLIP-2 is a multimodal model, which may not necessarily excel certain language methods in the field of NLP, thus we experiment with various language models (Devlin et al. 2019; Gao et al. 2021; AI@Meta 2024). Eventually, we employ LLaMA-3 to extract pooled features from D and C^f, denoted as $X^d = \{X_i^d\}_{i=1}^{N_d} \in \mathbb{R}^{N_d \times d}$ and $X^f = \{X_i^f\}_{i=1}^{L_v} \in \mathbb{R}^{L_v \times d}$, respectively. Next, we compute cosine similarity between X^d and X^f as the frame-level similarities $S^f \in \mathbb{R}^{N_d \times L_v}$:

$$S^f = \cos(X^d, X^f) = \frac{X^d \cdot X^f}{\|X^d\| \|X^f\|} \tag{1}$$

We are surprised to discover that this combination (image captioning and language-based frame scorer) can yield significant gains (see Tab. 5, row 5).

Span generator. After obtaining S^f, we design a span generator (SG) to dynamically produce candidates T^p. For clarity, we define $S_i^f \in \mathbb{R}^{L_v}$ as the cosine score between i-th debiased feature $X_i^d \in \mathbb{R}^d$ and X^f. Specifically, $S_{i,j}^f \in \mathbb{R}^1$ represents the score between X_i^d and j-th caption feature $X_j^f \in \mathbb{R}^d$.

Firstly, we compute the inverse cumulative histogram of S_i^f, with η bins. We then traverse these bins in reverse order to find the first bin containing at least κ moments, using its left endpoint value as the adaptive threshold γ. Next, we iterate through S_i^f in temporal order. If $S_{i,j}^f$ exceeds γ, the corresponding moment is marked as the starting moment. When the similarities of τ consecutive moments all fall below γ, we mark the final moment with a similarity exceeding γ as the ending moment. Finally, we repeat the above process to generate a set of candidate spans T^p from S^f:

$$T^p = \mathrm{SG}(S^f; \eta, \kappa, \tau) \tag{2}$$

where η , κ, and τ are hyperparameters. For a detailed case analysis, please see ***appendix***.

3.4 Choosing Relevant Spans

After getting T^p, previous methods (Huang et al. 2023a; Ren et al. 2023; Wattasseril et al. 2023) use multimodal models (Li et al. 2023a; Liu et al. 2023) to identify the spans most relevant to the query from T^p. However, visual redundancy in the video seriously impacts these methods' accuracy. Moreover, existing MLLMs (Maaz et al. 2023; Zhang et al. 2023) have yet to achieve satisfactory results in VMR, while they perform well in video captioning tasks. Therefore, we first employ Video-ChatGPT (Maaz et al. 2023) for video captioning to leverage the video comprehension capabilities of MLLMs. And then utilize a span scorer and post-processing to select the best matching spans $T \in \mathbb{R}^{N_t \times 2}$.

Video captioning. Fig. 4 (b) outlines the video captioning process, using Video-ChatGPT (Maaz et al. 2023) to create span-level captions $C^s \in \mathbb{R}^{N_p \times L_s}$. We devise the command: "`[Video caption] What is this video about? <Video>`". Here, `<Video>` represents the video segment corresponding to T^p.

Span scorer. We apply LLaMA-3 to extract the pooled features of C^s, denoted as $X^s \in \mathbb{R}^{N_p \times d}$. Subsequently, we compute the cosine similarity between X^s and X^d, resulting in a span-level similarity matrix $S^y \in \mathbb{R}^{N_p \times N_d}$, and then average it to obtain the span-level similarity $S^s \in \mathbb{R}^{N_p}$.

$$S^s = \mathrm{avg}[\cos(X^s, X^d)] \tag{3}$$

In our experiments, we observe that in spans containing multiple scenes, the frame-level similarities S^f sometimes exhibits several steep peaks with considerable intervals between them. In these cases, the span generator tends to construct shorter

Method	MLLM	Setting	QVHighlights test			QVHighlights val		
			R1@0.5	R1@0.7	mAP@avg	R1@0.5	R1@0.7	mAP@avg
VTimeLLM† (Huang et al. 2023a)	✓	FS	47.2	29.3	27.4	48.8	29.5	26.8
LLaViLo† (Ma et al. 2023)	✓	FS	48.6	29.7	27.9	49.0	30.4	28.9
Moment-DETR (Lei et al. 2021)		FS	52.9	33.0	30.7	54.2	33.4	31.1
UMT (Liu et al. 2022b)		FS	56.4	40.8	35.4	-	-	-
MomentDiff (Li et al. 2023b)		FS	-	-	-	57.8	39.2	35.3
CNM (Zheng et al. 2022a)		WS	14.1	4.0	-	-	-	-
CPL (Zheng et al. 2022b)		WS	30.8	10.8	-	-	-	-
CPI (Kong et al. 2023)		WS	32.3	11.8	-	-	-	-
(Liu et al. 2022a)		US	-	-	-	12.3	3.5	2.7
PZVMR (Wang et al. 2022a)		US	14.2	4.9	4.6	12.6	5.1	5.3
VideoLLaMA (Zhang et al. 2023)	✓	ZS	17.1	6.7	6.2	18.5	6.9	7.1
VideoChatGPT (Maaz et al. 2023)	✓	ZS	21.1	10.2	9.5	22.4	10.8	10.3
UniVTG (Lin et al. 2023)		ZS	25.2	9.0	10.9	-	-	-
(Diwan et al. 2023)	✓	ZS	-	-	-	48.3	31.0	28.0
(Wattasseril et al. 2023)†	✓	ZS	52.4	31.6	29.6	53.1	32.2	30.2
Moment-GPT (Ours)	✓	ZS	**58.3**	**37.7**	**35.0**	**58.9**	**38.6**	**35.9**

Table 1: Performance comparison with methods in different settings on QVHighlights. We denote FS for fully supervised, WS for weakly supervised (fine-tuning with both video and query), and US for unsupervised (fine-tuning with VMR-specific video). ZS indicates zero-shot (no VMR data fine-tuning required). "†" is our reproduced results.

candidates. To alleviate this issue, we calculate the span distance $E^s \in \mathbb{R}^{N_p}$ between the start and end moment and add it to the span-level score $S \in \mathbb{R}^{N_p}$, encouraging the retention of longer spans in the post-processed results.

$$S = (1 - \lambda) \cdot S^s + \lambda \cdot E^s \tag{4}$$

where λ represents the distance coefficient.

Post-processing. Finally, we sort the candidate spans T^p according to the score S and utilize NMS with an intersection-over-union (IoU) threshold σ to eliminate excessively overlapping spans, thereby obtaining the most suitable spans $T \in \mathbb{R}^{N_t \times 2}$:

$$T = \mathrm{NMS}(T^p, S; \sigma) \tag{5}$$

4 Experiments

4.1 Datasets and Metrics

Datasets. To evaluate our proposed method, we conduct experiments on three datasets with different topics: QVHighlights (Lei et al. 2021), Charades-STA (Gao et al. 2017), ActivityNet-Captions (Krishna et al. 2017). **QVHighlights** encompasses 10,310 queries and 10,148 YouTube videos with diverse topics, including daily vlogs, social news, et al. **Charades-STA**, derived from the Charades dataset (Sigurdsson et al. 2016), comprises 6,670 intricate indoor activity videos and 16,128 query-span pairs. **ActivityNet-Captions** originates from the ActivityNet dataset (Caba Heilbron et al. 2015), encompassing 19,811 outdoor activity videos and 71,957 pairs.

Metrics. To ensure fair comparison, we adhere to previous methods (Huang et al. 2023a; Luo et al. 2023), employing the following metrics for VMR: R1@n, mAP@avg, and mIoU. Expressly, R1@n signifies the percentage of test queries with at least one correctly localized span (IoU over n) in the top-1 outcomes. Likewise, mAP@avg denotes the average mAP over a set of IoU values [0.5: 0.05: 0.95]. mIoU is mean IoU.

4.2 Implementation Details

Following previous works (Huang et al. 2023a; Lei et al. 2021), we set the frame rates of videos from Charades-STA, ActivityNet-Captions, and QVHighlights to 1, 1, and 0.5, respectively. The employed MLLM models include LLaMA-3-8B, MiniGPT-v2-7B, and Video-ChatGPT based on Vicuna-7B-v1.1 (Zheng et al. 2024). To reduce the randomness of results, we configure the temperatures of LLaMA-3, MiniGPT-v2, and Video-ChatGPT to 0.3, 0.2, and 0.2, respectively. The number of histogram bins η is empirically fixed to 10. The hidden dimension d of LLaMA-3 is 4096. We set the number of debiased queries N_d to 3, the counting threshold κ to 7, the number of consecutive moments τ to 5, the distance coefficient λ to 0.2, and the IoU threshold σ in NMS to 0.9. All experiments are conducted on 1 NVIDIA A100 GPU.

4.3 Comparison With the State-of-the-Arts

To demonstrate the superiority of Moment-GPT, we first compare methods with different settings on QVHighlights (Tab. 1). Our approach demonstrates superior performance compared to the current SOTA zero-shot (ZS) method (Wattasseril et al. 2023), exhibiting at least **+5.4%** improvement in each metric. Moreover, it exceeds all FS (fully-supervised), WS (weakly-supervised), US (unsupervised), and MLLM-based methods. Tab. 2 reports the comparison on Charades-STA and ActivityNet-Captions. Our method achieves remarkable results, outperforming all ZS and US methods. Here, it outperforms the SOTA ZS method (Luo et al. 2023) by **+4.8%** and **+2.5%** in R1@0.3 on these two datasets, respectively.

Method	MLLM	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
GroundingGPT (Li et al. 2024)	✓	FS	-	29.6	11.9	-	-	-	-	-
VTimeLLM† (Huang et al. 2023a)	✓	FS	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeChat (Ren et al. 2023)	✓	FS	-	43.8	22.7	-	-	-	-	-
Moment-DETR† (Lei et al. 2021)		FS	62.1	48.2	25.3	42.3	52.6	32.5	15.3	37.8
CNM (Zheng et al. 2022a)		WS	50.0	36.2	14.2	34.2	51.3	30.3	11.4	33.9
CPL (Zheng et al. 2022b)		WS	56.0	38.1	20.3	37.8	52.4	30.9	12.0	32.6
(Huang et al. 2023b)		WS	59.2	44.2	22.1	39.4	54.8	32.9	-	36.4
PSVL (Nam et al. 2021)		US	45.2	30.9	14.2	30.9	45.1	29.8	15.73	30.2
(Gao et al. 2022)		US	45.3	19.8	7.9	-	45.8	25.9	12.1	-
(Liu et al. 2022a)		US	44.2	28.7	14.7	-	47.3	28.2	-	-
TimeChat (Ren et al. 2023)	✓	ZS	-	32.2	13.4	-	-	-	-	-
(Luo et al. 2023)†	✓	ZS	53.4	36.0	19.3	34.1	45.6	27.4	12.3	28.4
Moment-GPT (Ours)	✓	ZS	**58.2**	**38.4**	**21.6**	**36.5**	**48.1**	**31.1**	**14.9**	**30.8**

Table 2: Comparative evaluation on Charades-STA and ActivityNet-Captions.

Model	R1@0.5	R1@0.7	mIoU
LLaMA-2	37.7	20.9	35.7
Mistral-7B (Jiang et al. 2023)	37.2	21.1	35.4
LLaMA-3 (AI@Meta 2024)	**38.4**	**21.6**	**36.5**

Table 3: Different LLMs for query debiasing.

Model	R1@0.5	R1@0.7	mIoU
LLaVA (Liu et al. 2023)	37.1	20.3	35.2
MiniGPT-4 (Zhu et al. 2023)	37.8	20.7	35.9
MiniGPT-v2 (Chen et al. 2023)	**38.4**	**21.6**	**36.5**

Table 4: Different MLLMs for image captioning.

Additionally, it achieves competitive performance with FS and WS methods. We attribute the clear benefit of Moment-GPT over the aforementioned approaches to the query debiasing strategy and the effective utilization of MLLMs for generating and selecting spans.

4.4 Ablation Studies

To explore the effectiveness of each module, we conduct extensive ablation studies on Charades-STA.

Impact of LLMs. We first compare different LLMs (Touvron et al. 2023; Jiang et al. 2023; AI@Meta 2024) for reducing language bias under the same $N_d = 3$ condition (Tab. 3). Among them, LLaMA-3 demonstrates the highest performance, attributed to its extensive fine-tuning across diverse tasks and datasets, enabling superior text processing capabilities.

Image captioning and frame scorer. In Tab. 4, we present a comparison of different MLLMs (Zhu et al. 2023; Chen et al. 2023; Liu et al. 2023) for image captioning, where MiniGPT-v2 (Chen et al. 2023) attains the best results. Tab. 5 (left) assesses the influence of the frame scorer. In row 1,

Method	Frame scorer			Span scorer		
	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
BLIP-2	34.8	19.3	32.3	-	-	-
BLIP-2-T	36.9	19.7	34.2	-	-	-
BERT	37.3	20.8	35.4	37.5	20.4	35.8
SimCSE	38.0	21.1	35.9	37.7	21.0	36.1
LLaMA-3	**38.4**	**21.6**	**36.5**	**38.4**	**21.6**	**36.5**

Table 5: Effect of frame scorer (left) and span scorer (right).

BLIP-2 (Li et al. 2023a) is directly utilized to compute frame-level similarities between images and debiased queries. Row 2 integrates image captioning and employs BLIP-2's text encoder (BLIP-2-T) to obtain similarities. Contrasting rows 1 with 2 illustrates that deriving highly abstracted descriptions via MiniGPT-v2 effectively mitigates visual redundancy, thereby enhancing precision. The comparison between rows 2 and 3 reveals that language model (Devlin et al. 2019) is more effective than multimodal model in computing text similarity. Comparing rows 3-5 (Devlin et al. 2019; Gao et al. 2021), we finally choose LLaMA-3 for feature extraction in the frame scorer.

Effect of span generator. We scrutinize diverse strategies for span generation, as delineated in Tab. 6. For sliding window strategy (Ren et al. 2023), we set the window length to [10, 20, 30] with a sliding step of half the window length. Shot detection adheres to previous works (Diwan et al. 2023; Wattasseril et al. 2023), leveraging PySceneDetect[1] with consistent settings. The experimental findings show the superior performance of our span generator, demonstrating that our strategy of combining MiniGPT-v2 and frame scorer can effectively generate suitable spans.

Video captioning and span scorer. Tab. 7 evaluates the efficacy of various MLLMs for video captioning. Row 1 forego

[1]https://www.scenedetect.com/

Model	R1@0.5	R1@0.7	mIoU
Sliding Window	27.4	10.3	24.7
Shot Detection	32.1	11.7	30.9
Span Generator	**38.4**	**21.6**	**36.5**

Table 6: Different span generation methods.

Model	R1@0.5	R1@0.7	mIoU
None	30.2	12.4	28.7
MiniGPT-v2	32.4	14.3	30.7
VideoLLaMA	**38.9**	21.2	36.2
Video-ChatGPT	38.4	**21.6**	**36.5**

Table 7: Different MLLMs for video captioning.

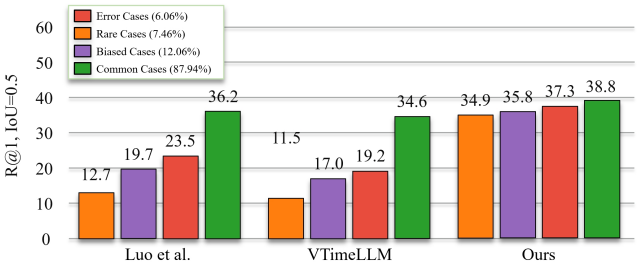


Figure 5: Analysis of the performance on biased and common cases on Charades-STA.

video captions and instead utilize the average frame-level similarity as span-level score. Row 2 extracts frames from video segments and feeds them into image-MLLM (MiniGPT-v2), resulting in poor outcomes. Conversely, row 4 leverages the video understanding capabilities inherent in video-MLLM (Video-ChatGPT), yielding superior performance. In Tab. 5 (right), we explore using different models to compute span-level similarity, with LLaMA-3 emerging as the most effective.

4.5 Further Analysis

Qualitative results. To qualitatively verify the effectiveness of our Moment-GPT, we visualize the localization outcomes of ground truth (GT), Luo et al. (Luo et al. 2023), VTimeLLM (Huang et al. 2023a), Moment-DETR (Lei et al. 2021), and Moment-GPT. Specifically, in Fig. 6 (top), our model detects the rare word "*kleenex*" in the query and corrects it to "*tissue*". In Fig. 6 (bottom), the rare word "*homerun*" is interpreted by the model as a more understandable term ("*hitting the ball out of the park*", "*complete circuit of the bases*"). In addition, grammatical errors and spelling errors in the query have been corrected. Moment-GPT achieves more accurate localization than other methods, especially in biased cases. The main reason is that previous methods rely only on the original query, which contains language bias. In contrast, our method could rectify incorrect queries, alleviate the bias derived from annotations, and leverage the video understanding capabilities of MLLMs to achieve more accurate localization.

Figure 6: Qualitative results on Charades-STA (top) and ActivityNet-Captions (bottom). We mark all biased and rewritten words in red.

Analysis on biased cases. In this paper, we categorize queries into different groups: `rare cases`, representing queries with at least one word (noun, verb) occurring less than 10 times; `error cases`, indicating queries containing spelling or grammatical errors; `biased cases`, which encompass the union of rare and error cases; and the remaining queries are termed `common cases`. As illustrated in Fig.5, we compare the performance across these four categories on Charades-STA. Our analysis yields the following insights: (1) Previous zero-shot method (Luo et al. 2023) and MLLM-based method (Huang et al. 2023a) fail to avoid language bias. (2) While VTimeLLM somewhat alleviates language bias through MLLM fine-tuning, its performance remains limited. (3) By optimizing queries using LLM, we successfully reduce language bias, resulting in improved results, particularly on biased cases.

5 Conclusion

This paper proposes **Moment-GPT**, a novel MLLM-based pipeline for zero-shot VMR, which utilizes frozen MLLMs for direct inference, avoiding fine-tuning. Furthermore, it reduces language bias in the original query and effectively leverages MLLMs' video comprehension abilities. Comprehensive experiments show that Moment-GPT considerably surpasses the SOTA MLLM-based and zero-shot methods on multiple datasets. Additionally, we deliver a comprehensive analysis that elaborates the design choices for each module. Finally, this work can be regarded as a MLLMs-driven multi-agent approach, offering valuable insights for VMR.

References

AI@Meta. 2024. Llama 3 Model Card.

Barrett, D. P.; Barbu, A.; Siddharth, N.; and Siskind, J. M. 2015. Saying what you're looking for: Linguistics meets video search. *IEEE TPAMI*, 38(10): 2069–2081.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; et al. 2023. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *github*.

Devlin, J.; et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Diwan, A.; et al. 2023. Zero-shot Video Moment Retrieval With Off-the-Shelf Models. In *Transfer Learning for Natural Language Processing Workshop*, 10–21. PMLR.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *CVPR*, 5267–5275.

Gao, J.; et al. 2022. Learning Video Moment Retrieval Without a Single Annotated Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1646–1657.

Gao, T.; et al. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP (1)*, 6894–6910. Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2023a. VTimeLLM: Empower LLM to Grasp Video Moments. ArXiv:2311.18445 [cs].

Huang, Y.; et al. 2023b. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18908–18918.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kong, S.; Li, L.; Zhang, B.; Wang, W.; Jiang, B.; Yan, C.; and Xu, C. 2023. Dynamic Contrastive Learning with Pseudo-samples Intervention for Weakly Supervised Joint Video MR and HD. In *ACM MM*, 538–546. Ottawa ON Canada.

Kong, W.; Hombaiah, S. A.; Zhang, M.; Mei, Q.; and Bendersky, M. 2024. PRewrite: Prompt Rewriting with Reinforcement Learning. *arXiv preprint arXiv:2401.08189*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *ICCV*, 706–715.

Lei, J.; et al. 2021. Detecting Moments and Highlights in Videos via Natural Language Queries. *NeurIPS*, 34: 11846–11858.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2023b. MomentDiff: Generative Video Moment Retrieval from Random to Real. *arXiv preprint arXiv:2307.02869*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; Huang, Z.; and Wang, T. 2024. GroundingGPT:Language Enhanced Multi-modal Grounding Model. ArXiv:2401.06071 [cs].

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; et al. 2020. Towards Debiasing Sentence Representations. In *ACL*, 5502–5515.

Liddy, E. D. 2001. Natural language processing.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. UniVTG: Towards Unified Video-Language Temporal Grounding. In *ICCV*, 2794–2804.

Linell, P. 2004. *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*, volume 36, 1683–1691. Issue: 2.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. ArXiv:2304.08485 [cs].

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022b. UMT: Unified Multi-modal Transformers for Joint Video Moment Retrieval and Highlight Detection. In *CVPR*, 3042–3051.

Luo, D.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2023. Zero-Shot Video Moment Retrieval from Frozen Vision-Language Models. ArXiv:2309.00661 [cs].

Lyu, Z.; and Zhang, Y. 2023. A novel temporal moment retrieval model for apron surveillance video. *Computers and Electrical Engineering*, 107: 108616.

Ma, K.; Zang, X.; Feng, Z.; Fang, H.; Ban, C.; et al. 2023. LLaViLo: Boosting Video Moment Retrieval via Adapter-Based Multimodal Modeling. In *ICCV*, 2798–2803.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. ArXiv:2306.05424 [cs].

Mekala, R. R.; et al. 2023. EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning. *arXiv preprint arXiv:2309.10687*.

Nam, J.; Ahn, D.; Kang, D.; Ha, S. J.; and Choi, J. 2021. Zero-shot natural language video localization. In *ICCV*, 1470–1479.

OpenAI. 2023. Introducing ChatGPT.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; and Clark, J. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2023. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. ArXiv:2312.02051 [cs].

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B. 2013. Translating video content to natural language descriptions. In *ICCV*, 433–440.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 510–526. Springer.

Sun, Y.; Xu, Y.; Xie, Z.; Shu, Y.; and Du, S. 2023. GPTSee: Enhancing Moment Retrieval and Highlight Detection via Description-Based Similarity Features. *IEEE Signal Processing Letters*. Publisher: IEEE.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Wang, G.; Wu, X.; Liu, Z.; and Yan, J. 2022a. Prompt-based Zero-shot Video Moment Retrieval. In *ACM MM*, 413–421. Lisboa Portugal: ACM. ISBN 978-1-4503-9203-7.

Wang, Y.; et al. 2022b. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. ArXiv:2212.03191 [cs].

Wattasseril, J. I.; Shekhar, S.; Döllner, J.; and Trapp, M. 2023. Zero-Shot Video Moment Retrieval Using BLIP-Based Models. In *Advances in Visual Computing*, volume 14361, 160–171. Cham.

Winograd, T. 1972. Understanding natural language. *Cognitive psychology*, 3(1): 1–191.

Wu, X.; Gao, C.; Lin, Z.; Wang, Z.; Han, J.; et al. 2022. RaP: Redundancy-aware Video-language Pre-training for Text-Video Retrieval. In *EMNLP*, 3036–3047.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; et al. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. ArXiv:2304.12244 [cs].

Xu, Y.; Sun, Y.; Zhai, B.; Jia, Y.; and Du, S. 2024a. MH-DETR: Video Moment and Highlight Detection with Cross-modal Transformer. In *IJCNN*, 1–8. IEEE.

Xu, Y.; Sun, Y.; Zhai, B.; Xie, Z.; Jia, Y.; and Du, S. 2024b. Multi-Modal Fusion and Query Refinement Network for Video Moment Retrieval and Highlight Detection. In *ICME*, 1–6. IEEE.

Yan, S.; Liu, M.; Wang, Y.; Liu, Y.; Chen, C.; and Liu, H. 2024. MLP: Motion Label Prior for Temporal Sentence Localization in Untrimmed 3D Human Motions. *arXiv preprint arXiv:2404.13657*.

Yu, X.; Jiang, C.; Dong, X.; Gan, T.; Yang, M.; et al. 2024. SHE-Net: Syntax-Hierarchy-Enhanced Text-Video Retrieval. *arXiv preprint arXiv:2404.14066*.

Zhang, H.; et al. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, volume 36, 3517–3525. Issue: 3.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, 15555–15564.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.