IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 September 2025, accepted 21 September 2025, date of publication 23 September 2025,
date of current version 29 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3613635

RESEARCH ARTICLE

DBCON: Dual Bias Control in Zero-Shot Video Moment Retrieval

MINGYU JEON[1], (Member, IEEE), MINUK MA[2], AND JUNYEONG KIM[1], (Member, IEEE)
[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Computer Science, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Chung-Ang University) under Grant RS-2021-II211341; and in part by IITP Grant funded by Korea Government (MSIT) (Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) under Grant RS-2022-II220184.

ABSTRACT Zero-shot Video Moment Retrieval (ZVMR) is a task that localizes temporal moments within a video using only natural language queries, without requiring large-scale datasets or extensive training resources. While previous ZVMR studies have focused on aligning high-quality video and query representations, they have faced performance limitations by overlooking the 'Dual Bias' problem, which arises from the excessive specificity of queries and the generality of video captions. To address this challenge, we propose the DBCON (Dual Bias CONtrol) framework, which simultaneously controls caption specificity and query generalization. DBCON employs 'Query-Aware Caption Refinement' to induce captions that explicitly reflect the core information of the query, and 'Query Debiasing' to rephrase the user query into a more universal expression via an LLM. Our experiments show that DBCON effectively mitigates this dual bias, achieving up to 1.73% higher retrieval accuracy on mAP@avg compared to the state-of-the-art model on the QVHighlights test benchmark. This study is significant as it presents a new direction for developing scalable and efficient video-language retrieval systems.

INDEX TERMS Video moment retrieval, zero-shot Learning, multimodal learning, video understanding, bias control, specificity mismatch, natural language processing.

I. INTRODUCTION

The recent explosion of digital content has heightened the need for technologies that can accurately and efficiently retrieve specific moments from vast video archives. Video Moment Retrieval (VMR), a key technology addressing this demand, aims to identify specific temporal segments within a video based on a natural language query. However, most existing VMR methods have a significant limitation: they rely on large-scale annotated datasets and substantial training resources. This dependency hinders scalability to new domains and severely reduces training efficiency in resource-constrained environments.

To overcome these limitations, zero-shot Video Moment Retrieval (ZVMR) has emerged as a promising alternative. ZVMR utilizes only natural language queries and raw video content, eliminating the need for task-specific labels or further training. This approach demonstrates strong generalization across diverse domains and is highly effective in data-scarce scenarios.

Nevertheless, ZVMR still faces two critical challenges, which we define as a form of Dual Bias that creates a specificity mismatch: (1) Semantic Distortion Bias in Queries: As illustrated in Figure 1, user queries containing excessive specificity (e.g., ''A fluffy white Persian''), subjective modifiers, or grammatical errors can create a mismatch with the generalized representation space of pre-trained models, acting as a Semantic Distortion Bias. This has a negative effect, unfairly lowering the similarity scores of relevant video regions and thus hindering a precise semantic link with the video content. (2) Representation Deficiency Bias in Captions: Conversely, generic captions

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

generated by VLMs (e.g., "A cute cat") suffer from a Representation Deficiency Bias, as they lack the information needed to capture the specific intent of a query (e.g., "A fluffy white Persian cat is yawning"). These captions lack semantic richness and fail to reflect the query's specific intent, consequently causing the model to miss relevant moments or incorrectly retrieve irrelevant segments.

In this paper, we propose DBCON (Dual Bias CONtrol), a novel framework designed to resolve this "specificity mismatch" problem by simultaneously managing both the Semantic Distortion Bias from the query and the Representation Deficiency Bias from the caption. DBCON addresses this issue through two core modules. The first module, Query-Aware Caption Refinement, uses the original query as an explicit condition to generate more relevant captions that align with the query's intent. In parallel, the second module, Query Debiasing, employs a large language model (LLM) to rephrase the original user query into more universal terms. This process alleviates the bias caused by the linguistic specificity or uncommon expressions in the user query, helping the model focus on the core semantics.

The main contributions of this work are as follows:

- This work redefines the "specificity mismatch" in ZVMR, a key cause of performance degradation, from the perspective of a dual bias between queries and captions.
- We propose DBCON, a novel framework that effectively bridges the semantic gap through a complementary strategy of Query-Aware Caption Refinement and Query Debiasing.
- The proposed method achieves new state-of-the-art (SOTA) performance on the QVHighlights zero-shot VMR benchmark; this training-free achievement demonstrates its superiority, efficiency, and high scalability.

We believe that DBCON makes a significant contribution toward developing scalable, domain-adaptable, and efficient video-language retrieval systems.

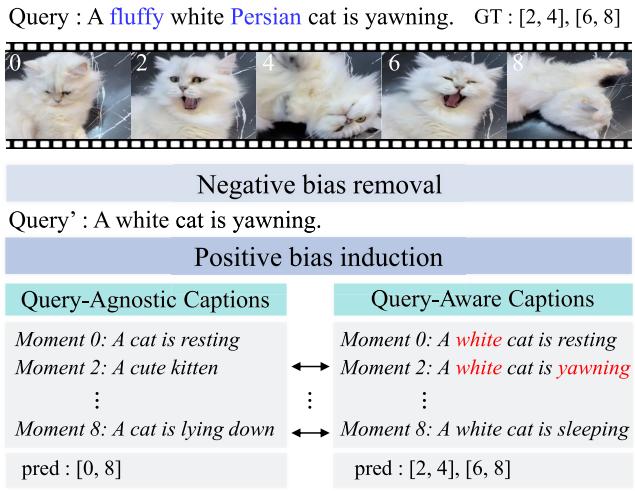


FIGURE 1. Example of negative bias removal and positive bias induction for the query "A fluffy white Persian cat is yawning." The debiased query is generalized, and query-aware captions better reflect the query's intent. Compared to conventional captions, our method improves semantic alignment and enables more accurate segment prediction.

II. RELATED WORK

A. ZERO-SHOT VIDEO MOMENT RETRIEVAL

Zero-shot Video Moment Retrieval (ZVMR) aims to locate a video segment corresponding to a given query without task-specific training. Early ZVMR studies [6], [10], [11], [31] primarily relied on off-the-shelf pipelines that leverage large-scale, pre-trained Vision-Language Models (VLMs). This approach typically calculates the similarity between each video frame and the query to find semantically matching segments, achieving reasonable performance on benchmarks like QVHighlights [1]. However, this method's reliance on coarse, frame-level matching imposes clear limitations on its ability to infer fine-grained temporal relationships or understand complex contexts.

To improve upon these limitations, methods like UniVTG [7] emerged, which sought to enhance model generalization by unifying several related tasks, such as temporal grounding. While UniVTG demonstrated a better temporal understanding through its unified training, it still struggled to capture complex and subtle temporal relations. More recently, a new paradigm has emerged that incorporates Large Language Models (LLMs) into ZVMR. The current state-of-the-art method, Moment-GPT [12], exemplifies this by using an LLM to rephrase an original query into diverse expressions and then using a video-language model like Video-ChatGPT [20] to rank candidate moments.

Although Moment-GPT [12] secures representational diversity through query rephrasing, this strategy only solves half of the problem. The rephrased queries can subtly deviate from the original intent or create new alignment issues with the visual content of the video frames. In other words, improving the query side alone is insufficient to overcome the semantic generality inherent in the video-side representations. Our framework, DBCON, recognizes this limitation and addresses the problem with a dual approach that not only refines the query but also actively enhances the video captions to align with the query.

B. QUERY-FOCUSED VIDEO PROCESSING

Research leveraging queries to enhance the accuracy of video understanding has been actively pursued for years. One major stream of work focuses on generating frame- or segment-level captions to strengthen video-text alignment itself. These methods enrich the semantic representation of the video, but they primarily focus on generating high-quality, generic captions independent of the user's immediate query [19], [21], [25], [26].

Another line of research involves selecting or summarizing specific parts of a video based on the user's query. Techniques like query-focused video summarization can effectively

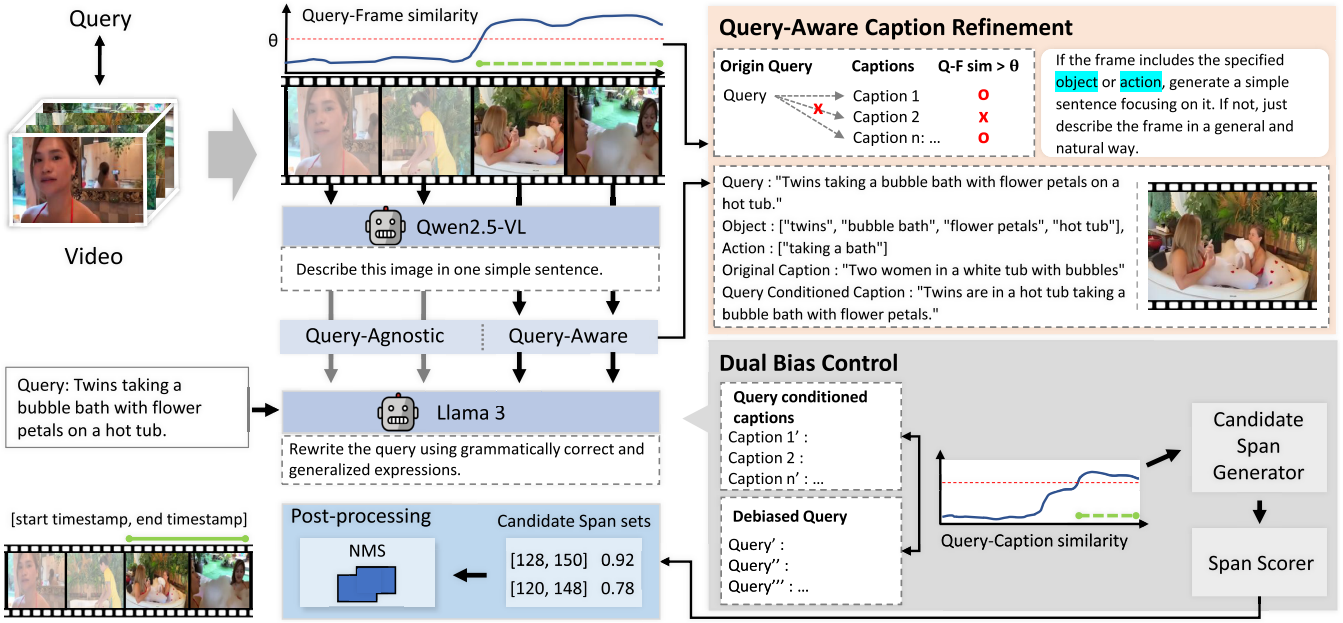


FIGURE 2. **Overview of the proposed DBCON framework. The figure illustrates the complete pipeline of DBCON. Given a raw video and a user query, we first identify initial candidate spans via query-frame similarity. The core of DBCON involves two parallel processes: (1) The Query-Aware Caption Refinement module (top right) parses the original query and uses a VLM (Qwen2.5-VL) to generate detailed, query-aware captions for the candidate spans. (2) Simultaneously, the Query Debiasing module uses an LLM (Llama 3) to generalize the original query into a more robust, debiased version. Finally, the Dual Bias Control module computes the similarity between the debiased query and the query-aware captions to re-evaluate and score the candidate spans, which are then finalized using NMS post-processing. This dual approach ensures that a robust, generalized query is matched against a highly specific, context-rich video representation, effectively resolving the specificity mismatch.**

identify relevant scenes [14], [16], [17], but they are not suited for localizing the precise temporal segments that VMR requires. Meanwhile, some studies have attempted to control bias at the feature level by modulating internal attention mechanisms or filtering representations [13], [15], [18].

Prior works have thus concentrated on unidirectional improvements, such as enhancing caption quality or using the query for summarization and feature control. The novelty of our DBCON lies in its explicit and joint application of Query Debiasing and Query-Aware Caption Refinement within a ZVMR framework to solve the problem from both directions. This is a new attempt that goes beyond simply generating better captions; it aims to create query-aware captions while simultaneously mitigating the bias of the query itself to fundamentally resolve the ''specificity mismatch'' problem.

III. PROPOSED METHOD

A. OVERVIEW

In this paper, we propose the **DBCON (Dual Bias CONtrol)** framework to address the critical problem of ''specificity mismatch'' in Zero-shot Video Moment Retrieval (ZVMR). DBCON operates through a three-stage pipeline designed to progressively enhance the semantic alignment between the query and video content: (1) Initial Candidate Span Identification, (2) Dual Bias Control, and (3) Span Re-evaluation and Finalization. An overview of the entire process is illustrated in Figure 2.

B. INITIAL CANDIDATE SPAN IDENTIFICATION

The first stage of DBCON rapidly identifies a set of initial, high-relevance candidate spans from the original video and query. The inputs are an untrimmed video, defined as $V = \{f_i\}_{i=1}^{L_v}$, where f_i denotes the i-th frame and L_v is the total number of frames in the video, along with the original text query Q. First, we use a pre-trained vision-language model (CLIP ViT-B/32) to extract feature vectors from each video frame and the original query. These are denoted as frame features $E_v = \{e_i^f\}_{i=1}^{L_v}$ and query features E_{q_o}, respectively. We then compute an initial relevance score by calculating the cosine similarity $S_{init}(i) = \text{sim}(e_i^f, E_{q_o})$ for each frame. Finally, continuous frames with scores exceeding a predefined threshold, θ_{init}, are grouped to form the set of initial candidate spans, T_{init}. This stage serves to reduce computational complexity and focus subsequent processing on the most relevant video regions.

C. DUAL BIAS CONTROL

This stage is the core of DBCON, where we actively refine both the query-side and video-side representations to resolve the ''specificity mismatch'', as illustrated in Figure 3. It consists of two sub-modules: Query-Aware Caption Refinement and Query Debiasing.

1) QUERY-AWARE CAPTION REFINEMENT

Automatically generated captions are often too generic (e.g., ''a man is present'') to reflect the specific intent of

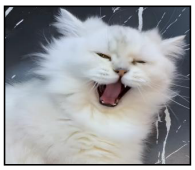


FIGURE 3. An illustration of the semantic alignment challenge solved by DBCON. A generic, query-agnostic caption for the yawning cat only describes the physical state ('with its mouth open'), creating a semantic mismatch with the rephrased query's intent ('yawn'). In contrast, our query-aware caption, guided by the raw query, explicitly includes the action 'yawn,' resulting in a strong semantic alignment and enabling accurate retrieval.

a query. To solve this, we actively refine the captions to be query-aware, rather than passively describing the video content. This can be seen as a form of positive bias injection, enriching the video-side representation with query-specific context.

Our refinement process begins by parsing the original query Q to extract its thematic keywords. To do this, we employ a lightweight syntactic parser that uses Part-of-Speech (POS) tagging and dependency analysis to identify the primary nouns (as objects) and main verbs (as actions). These extracted keywords are then inserted into a dynamic conditional prompt template for Qwen2.5-VL [23].

The prompt instructs the VLM to operate conditionally: if the specified objects or actions are visually present in a frame, the VLM must generate a caption that emphasizes them. If they are not present, it defaults to a general, objective description. This process is applied to frames within the initial candidate spans, T_{init}, yielding a set of query-aware captions, denoted as C_{ref}. Unlike generic, query-agnostic descriptions, these captions are specifically tailored to the user's intent, creating a more detailed and context-rich representation of the relevant video content.

2) QUERY DEBIASING

User queries often present challenges such as grammatical errors, subjective language, or excessive specificity (e.g., "a fluffy white Persian"). These issues introduce a negative bias that can hinder model generalization. To address this, we use Llama3 [24] to rewrite the original query Q into a cleaner and more general expression.

Our prompt guides the LLM to analyze the query by separating its core components, like primary objects and actions, from specific or subjective modifiers. The LLM preserves the query's core intent while making two key changes. First, it corrects any grammatical errors. Second, it generalizes overly specific modifiers into more universal terms; for example, "fluffy white Persian" is generalized to the more universal phrase "white cat with long fur."

This process yields a refined query, Q', that is both syntactically sound and semantically robust. By preserving the original intent in a more generalized form, Q' can be aligned with a broader range of visual representations. We term this step "negative bias removal", as it is crucial for improving alignment success in the subsequent stage.

D. SPAN RE-EVALUATION AND FINALIZATION

In the final stage, we use the refined query Q' produced by the dual bias control module. We also utilize the full set of frame captions, which includes the query-aware captions C_{ref}. These inputs are used to evaluate the candidate spans and generate the final result.

1) CANDIDATE SPAN GENERATION

We compute a refined relevance score, $S_{ref}(i)$, by calculating the cosine similarity between the caption for every frame (using C_{ref} for frames in candidate spans and generic captions otherwise) and the debiased query Q'. Then, using an inverse cumulative histogram based on this new score distribution, we select second-stage candidate spans from the top-K bins. This method allows for a more flexible selection of candidates based on their relative importance within the distribution, rather than relying on a fixed absolute threshold.

2) SPAN SCORING

The final score s_k for each candidate span T_k is computed as a weighted combination of its average semantic relevance and temporal continuity:

$$s_k = \alpha \cdot S_{avg}(T_k) + (1 - \alpha) \cdot D(T_k)$$

Here, $S_{avg}(T_k)$ is the average of the refined relevance scores (S_{ref}) of all frames within the candidate span T_k. This term represents how semantically close the span is to the query's content. The term $D(T_k)$ is a continuity adjustment term that rewards longer spans. Experimentally, we observed that in long videos, relevant scenes can appear with temporal gaps, causing similarity scores to exhibit multiple steep peaks. This often leads the model to predict short, fragmented spans centered on each peak. To alleviate this issue, $D(T_k)$ uses the normalized length of the candidate span, calculated by dividing its length by the total length of all candidate spans. Thus, $D(T_k)$ acts as a length-aware regularization factor. The hyperparameter α controls the balance between semantic relevance (S_{avg}) and temporal continuity ($D(T_k)$).

TABLE 1. Performance comparison on QVHighlights test and val sets. "MLLM" indicates the use of a multi-modal large language model, and "FS/WS/US/ZS" denote fully-supervised, weakly-supervised, unsupervised, and zero-shot settings, respectively.

Method	MLLM	Setting	QVHighlights test				QVHighlights val			
			R1@0.5	R1@0.7	mAP@0.5	mAP@avg	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
VTimeLLM [5]	✓	FS	47.2	29.3	47.3	27.4	48.8	29.5	49.3	26.8
LLaViLo [2]	✓	FS	48.6	29.7	48.7	27.9	49.0	30.4	49.4	28.9
Moment-DETR [1]	–	FS	52.9	33.0	54.8	30.7	54.2	33.4	55.4	31.1
UMT [35]	–	FS	56.4	40.8	53.1	35.4	-	-	-	-
MomentDiff [36]	–	FS	-	-	-	-	57.8	39.2	54.6	35.3
CNM [27]	–	WS	14.1	4.0	11.8	–	–	–	–	–
CPL [13]	–	WS	30.8	10.8	22.8	–	–	–	–	–
CPI [9]	–	WS	32.3	11.8	23.7	–	–	–	–	–
Liu et al. [8]	–	US	–	–	–	–	12.3	3.5	10.4	2.7
PZVMR [10]	–	US	14.2	4.9	15.7	4.6	12.6	5.1	16.2	5.3
VideoLLaMa [34]	✓	ZS	17.1	6.7	18.2	6.2	18.5	6.9	17.8	7.1
VideoChatGPT [20]	✓	ZS	21.1	10.2	22.8	9.5	22.4	10.8	21.9	10.3
UniVTG [7]		ZS	25.2	9.0	27.4	10.9	–	–	–	–
Diwan [6]	✓	ZS	–	–	–	–	48.3	31.0	47.3	28.0
Moment-GPT [12]	✓	ZS	58.3	37.7	55.1	35.0	58.9	38.6	55.7	35.9
DBCON (ours)	✓	ZS	**58.91**	**38.39**	**56.76**	**36.73**	**60.65**	**41.03**	**58.12**	**37.53**

3) POST-PROCESSING

Finally, to further enhance continuity, we merge adjacent candidate spans if they are separated by up to P consecutive frames with low similarity. However, to prevent erroneous merging across highly irrelevant segments, this process is suppressed if the frames within the gap fall into the bottom $\theta\%$ of the overall similarity distribution. The scored spans then undergo Non-Maximum Suppression (NMS) to eliminate overlaps, yielding the final set of predicted segments, T_{final}. The specific values for each hyperparameter are detailed in the Implementation Details section.

IV. EXPERIMENTS

This section details the experiments conducted to validate the effectiveness of the proposed DBCON framework. We present the experimental setup, a performance comparison against state-of-the-art (SOTA) methods, an ablation study analyzing the contributions of our core components, and an in-depth analysis to demonstrate the robustness of our method.

A. DATASETS AND METRICS

1) DATASETS

Our performance evaluation was conducted on three widely used benchmarks: QVHighlights [1], Charades-STA [3], and ActivityNet-Captions [4].

QVHighlights is a large-scale benchmark comprising 10,148 YouTube videos with an average length of 150 seconds, providing a realistic retrieval scenario. The dataset features an average query length of 11.3 words and an average target moment duration of 24.6 seconds. It is divided into 7,218 queries for training, 1,550 for validation, and 1,542 for testing. For fair comparison, the ground truth for the test set is hidden, and performance is measured via submission to an official evaluation server.

To verify our method's generalization performance, we use Charades-STA, which focuses on daily life activities. It consists of 9,848 videos with an average length of 30.6 seconds and contains 16,128 temporal annotations, with an average moment length of 8.1 seconds. ActivityNet-Captions is a benchmark of 20,000 untrimmed videos totaling approximately 849 hours, paired with about 100,000 captions (an average of 3.6 moments per video), and is widely used for dense-event captioning and moment retrieval.

2) METRICS

Model performance is evaluated using standard VMR metrics, primarily Recall@1 (R@1) and mean Average Precision (mAP). R@1 measures the proportion of queries for which the top-ranked predicted segment exceeds a certain Intersection over Union (IoU) threshold. The mAP metric quantifies the average precision across various recall levels at a given IoU threshold.

Following the standard protocol for each benchmark, we report these metrics at different IoU thresholds. For QVHighlights, our primary metrics are R@1@0.5, R@1@0.7, mAP@0.5, and the average mAP across multiple thresholds (mAP@avg). For Charades-STA and ActivityNet-Captions, we report R@1 at IoU thresholds of 0.3, 0.5, and 0.7, along with the mean IoU (mIoU).

All evaluations were performed under a zero-shot setting, meaning no training was conducted on the evaluation datasets.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

To objectively evaluate DBCON's performance, we compared it against recent VMR models in a zero-shot setting across all three benchmark datasets (see Table 1, 2).

On QVHighlights [1], DBCON demonstrated its effectiveness by outperforming all competitors, including the

TABLE 2. Performance comparison on Charades-STA and ActivityNet-Captions under various supervision settings.

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 [32]	✓	FS	–	29.6	11.9	–	–	–	–	–
VTimeLLM [5]	✓	FS	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeChat [33]	✓	FS	–	43.8	22.7	–	–	–	–	–
Moment-DETR [1]	–	FS	62.1	48.2	25.3	42.3	52.6	32.5	15.3	37.8
CNM [27]	–	WS	50.0	36.2	14.2	34.2	51.3	30.3	11.4	33.9
CPL [13]	–	WS	56.0	38.1	20.3	37.8	52.4	30.9	12.0	32.6
Huang et al. [28]	–	WS	59.2	44.2	22.1	39.4	54.8	32.9	–	36.4
PSVL [29]	–	US	45.2	30.9	14.2	30.9	45.1	29.8	15.73	30.2
Gao et al. [30]	–	US	45.3	19.8	7.9	–	45.8	25.9	12.1	–
Liu et al. [8]	–	US	44.2	28.7	14.7	–	47.3	28.2	–	–
TimeChat [33]	✓	ZS	–	32.2	13.4	–	–	–	–	–
Luo et al. [31]	✓	ZS	53.4	36.0	19.3	34.1	45.6	27.4	12.3	28.4
Moment-GPT [12]	✓	ZS	58.2	38.4	21.6	36.5	48.1	31.1	14.9	30.8
DBCON (Ours)	✓	ZS	**58.8**	**39.1**	**22.1**	**37.2**	**49.2**	**32.9**	**15.4**	**32.0**

TABLE 3. Ablation study showing the effect of the original query (Q), debiased query (Q'), General Captions (GC), and Query-Aware Captions (QC) on performance in the QVHighlights validation set.

Config.	Q	Q'	GC	QC	R1@0.5	R1@0.7
(a)	✓	–	✓	–	57.94	38.84
(b)	✓	–	–	✓	58.19	39.03
(c)	–	✓	✓	–	60.0	40.32
(d)	–	✓	–	✓	**60.65**	**41.03**

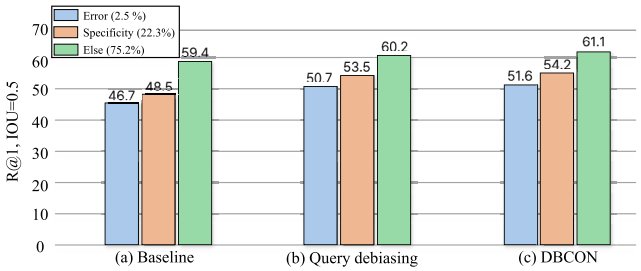


FIGURE 4. Fine-grained analysis of DBCON's performance (R1@0.5) on different query subsets. The chart compares the performance of the Baseline, a model with only Query Debiasing, and the full DBCON framework. The results clearly demonstrate that the performance gains are most pronounced on the 'Error' and 'Specificity' sets, validating that our dual bias control approach is particularly effective for handling challenging queries with syntactic errors or high semantic specificity.

previous SOTA model Moment-GPT [12], across all zero-shot evaluation metrics. This strong performance was consistent on other benchmarks as well. On both the Charades-STA [3] and ActivityNet-Captions [4] datasets, DBCON showed competitive performance compared to other zero-shot approaches. This indicates that our dual bias control strategy is not limited to a specific dataset but is robust for various types of videos and queries. In particular, it shows that our approach of directly improving the semantic alignment of queries and captions is effective for securing generalization performance across diverse domains.

C. ABLATION STUDY

To quantify the contribution of each component in the DBCON framework, we performed an ablation study on the QVHighlights validation set (Table 3). We denote the original user query as Q and its debiased form, produced by our Query Debiasing module, as Q'. Frame-level captions generated without any query conditioning are referred to as General Captions (GC), whereas captions conditioned on a query, either Q or Q', are called Query-Aware Captions (QC, i.e., C_{ref}). Four experimental configurations were evaluated: (a) the baseline, which pairs Q with GC; (b) a setting isolating caption conditioning by combining Q with QC; (c) a setting isolating debiasing by combining Q' with GC; and (d) the full DBCON model, which combines both innovations by pairing Q' with QC. The results show that debiasing the query yields a larger standalone gain than caption conditioning, and that integrating the two techniques delivers the highest Recall@0.5 and Recall@0.7, confirming their complementary effects.

The results clearly show the impact of each module. Compared to the baseline (a), applying only Query-Aware Caption Refinement (b) yielded a slight performance increase. In contrast, isolating Query Debiasing (c) led to a much more significant improvement, demonstrating it is the more dominant component for standalone performance gain. Finally, the full DBCON model (d), which integrates both techniques, achieved the highest scores. This confirms that while query debiasing is more impactful on its own, the two components are complementary and create a synergistic effect when combined. Furthermore, we conducted additional ablation studies to validate our choice of key

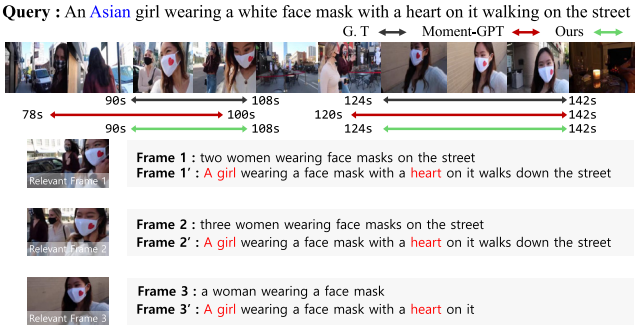


FIGURE 5. Qualitative results on the QVHighlights validation set. The figure illustrates how bias control is applied during query-aware caption generation. Specifically, the removal of "Asian" demonstrates negative bias removal, while the inclusion of specific attributes such as "girl" and "heart" reflects positive bias injection based on the original query intent. Frame' indicates a query-aware captioned frame.

hyperparameters, with detailed results presented in Tables 4 through 9.

D. FINE-GRAINED ANALYSIS OF QUERY DEBIASING

As shown in Figure 4, to precisely isolate the contributions of our Dual Bias Control framework, we conducted a fine-grained analysis. For this, we partitioned the QVHighlights validation set into three distinct categories based on query characteristics: a **Grammatical Error Set** (2.5% of queries) containing syntactic flaws; a **High Specificity Set** (22.3%) comprising queries with rare terminology or a high density of attributes; and the remaining **Standard Query Set** (75.2%). On these subsets, we compared the R@1@0.5 performance of three model configurations: (a) a Baseline, using the original query with query-agnostic captions; (b) Query Debiasing only, pairing the debiased query with query-agnostic captions; and (c) our full DBCON framework, which uses both the debiased query and query-aware captions.

The analysis revealed that the performance gains from our method were most pronounced on the two challenging sets. On the Grammatical Error Set, applying Query Debiasing alone improved performance by a significant 4.0% over the baseline (from 46.7% to 50.7%), with the full DBCON framework further increasing the score to 51.6%. The synergy of our dual-bias approach was most apparent on the High Specificity Set, where the full DBCON framework achieved a final score of 54.2%, marking a total uplift of 5.7% from the baseline of 48.5%. This demonstrates that the combination of a cleaned query and a query-aware caption is crucial for resolving complex requests. In contrast, for the Standard Query Set, where queries are already well-formed, the improvement was more modest, with the full framework providing a 1.7% gain over the baseline (from 59.4% to 61.1%). This fine-grained analysis provides robust evidence that DBCON's primary contribution lies in systematically and robustly handling difficult, real-world query cases.

E. QUALITATIVE ANALYSIS

In addition to quantitative evaluation, we performed a qualitative analysis to provide an intuitive understanding of how DBCON operates. In the example from Figure 5, the query is "An Asian girl wearing a white face mask with a heart on it walking on the street." A generic caption generated by conventional methods misses the key detail ("face mask with a heart on it"), leading to failure in accurately localizing the moment. In contrast, DBCON's Query-Aware Caption Refinement module uses the core information from the original query to generate a highly detailed and accurate caption, such as "A girl wearing a face mask with a heart on it walks down the street". This enables the successful retrieval of the correct video segment that was previously missed. These examples clearly illustrate how DBCON effectively bridges the semantic gap between the user's intent and the video's content.

F. IMPLEMENTATION DETAILS

All experiments were conducted on a system with 4 NVIDIA A6000 GPUs. Frame-level captions were generated using the Qwen2.5-VL-7B [23] model on frames sampled at 0.5 fps. We used the CLIP [22] ViT-B/32 model for feature extraction from both frames and queries. To focus on the most relevant content, we then selected only the captions corresponding to frames whose initial similarity to the query ranks within the top 10% for further processing (see Table 4). For our core modules, Query Debiasing was performed using the Llama3 [24] model to generate 5 rewritten queries, each designed to preserve the core semantic meaning of the original. We found that this number of queries yields the best performance empirically (see Table 5).

We selected Qwen2.5-VL and Llama 3 because their combination demonstrated the optimal synergy in executing our framework's specific instructions. To validate the general effectiveness of our approach and demonstrate that it is not reliant on a single model pair, we conducted an ablation study with other representative open-source models, including LLaVA [37], BLIP-2 [38], and Mistral 7B [39]. As shown in Table 10, while other combinations maintain competitive performance, the Qwen-Llama pair was most successful at executing both the sophisticated conditional captioning of 'Query-Aware Caption Refinement' and the semantic generalization of 'Query Debiasing'. This validates our model choice while also demonstrating the inherent robustness of our dual-bias control method.

Notably, our DBCON framework is designed to be model-agnostic, meaning it is not dependent on these specific models. This architectural flexibility allows for easy integration of more powerful VLMs or LLMs as they become available, enabling continuous improvements to the overall system performance simply by replacing the corresponding modules.

In the Candidate Span Generation stage, we employed an inverse cumulative histogram with 10 bins and selected candidates from the top 8. To enhance prediction continuity,

TABLE 4. Performance variation according to the frame ratio for query-aware captioning on Charades-STA.

Top-K%	R1@0.3	R1@0.5	mIoU
5%	57.9	38.0	36.3
10%	**58.8**	**39.1**	**37.2**
20%	58.6	38.8	37.0
30%	58.3	38.4	36.7

TABLE 5. Performance variation according to the number of rewritten queries on QVHighlights val set. N_q denotes the number of queries.

N_q	R1@0.5	R1@0.7
1	58.72	39.15
3	60.18	40.38
5	**60.65**	**41.03**
7	59.91	40.12
9	58.32	39.61

TABLE 6. Performance variation according to the maximum merging gap (P) on Charades-STA.

P	R1@0.3	R1@0.5	mIoU
2	57.2	37.4	35.8
4	58.1	38.5	36.7
6	**58.8**	**39.1**	**37.2**
8	58.5	38.7	36.9

TABLE 7. Performance variation according to the merging suppression threshold θ on Charades-STA.

θ	R1@0.3	R1@0.5	mIoU
0%	57.8	37.9	36.1
10%	58.3	38.6	36.7
20%	**58.8**	**39.1**	**37.2**
30%	58.6	38.8	37.0

TABLE 8. Performance variation according to the weight α on QVHighlights val set.

Weight α	R1@0.5	R1@0.7
0.1	57.88	38.44
0.3	59.45	40.21
0.5	**60.65**	**41.03**
0.7	60.23	40.79
0.9	58.91	39.62

TABLE 9. Ablation on the NMS IoU threshold (θ_{NMS}) on the Charades-STA dataset.

θ_{NMS}	R1@0.3	R1@0.5	mIoU
0.8	58.1	38.4	36.5
0.85	58.4	38.6	36.9
0.9	**58.8**	**39.1**	**37.2**
0.95	58.6	38.9	37.0

candidate spans separated by a gap of up to P=6 consecutive low-similarity frames are merged (see Table 6); however, this merging is suppressed if the frames within the gap fall into the bottom $\theta = 20\%$ of the similarity distribution (see Table 7). The weight α in the final Span Scorer was set to 0.5 (see Table 8), and the IoU threshold for Non-Maximum Suppression (NMS) was set to 0.9 (see Table 9). Unless otherwise specified, all key thresholds were empirically determined based on performance on a validation set. Furthermore, our experiments confirmed that the framework is not overly sensitive to the choice of these hyperparameters.

TABLE 10. Ablation study on the choice of VLM and LLM on the QVHighlights validation set. Our selected models are shown in Bold.

VLM	LLM	R1@0.5	R1@0.7
Qwen2.5-VL	**Llama 3 (Ours)**	**60.65**	**41.03**
LLaVA	Llama 3	58.21	39.75
BLIP-2	Llama 3	57.07	37.51
Qwen2.5-VL	Mistral 7B	60.15	39.88

V. LIMITATIONS

While the proposed DBCON framework is effective in resolving the core challenge of 'Dual Bias', it entails the following two primary limitations. First is the online inference cost. DBCON performs 'online' computations such as caption generation and query rewriting, which introduces latency in real-world applications and poses challenges for deploying the system in environments requiring immediate responsiveness. Second, the framework's performance is highly dependent on the quality of external generative models. A single error from these models, which are responsible for caption generation and query rewriting, can lead to incorrect retrievals and degrade the overall reliability and robustness of the system.

VI. FUTURE WORK

Based on the limitations of the current framework, we propose several practical directions for future work that can enhance the system's efficiency and robustness without requiring new training. First, the current high-cost pipeline, which involves generating and comparing text captions via a VLM, can be optimized. To this end, we propose a method that bypasses the text generation process and utilizes the query directly at the feature level. Specifically, this involves directly comparing the query feature, obtained by embedding the user query with a model like CLIP, against video frame features within the feature space. This approach completely bypasses the VLM's text generation process, thereby significantly improving the system's responsiveness and practicality. Furthermore, to mitigate the dependency on external model quality, the system's robustness can be secured by introducing a validation mechanism at the inference stage that utilizes multiple query candidates. This method leverages the multiple query candidates already generated by the framework to cross-validate whether a candidate video segment consistently scores high across various query phrasings. This would allow the system to

effectively filter incorrect retrievals and improve stability against model errors without requiring additional training.

VII. CONCLUSION

In this study, we defined the core challenge of Zero-shot Video Moment Retrieval (ZVMR) as a 'Dual Bias' problem, which arises from the excessive specificity of queries and the generality of video captions, and proposed the DBCON framework to address it. Through the complementary action of 'Query Debiasing,' which controls bias on the query side, and 'Query-Aware Caption Refinement,' which compensates for bias on the caption side, DBCON effectively achieves semantic alignment between user intent and video content.

Experiments have demonstrated that our framework successfully mitigates this dual bias, achieving state-of-the-art (SOTA) performance across major benchmarks. DBCON presents a novel dual-control approach that simultaneously manages representations on both the query and video sides, offering a significant direction for developing more robust and adaptable video-language retrieval systems in the future.

REFERENCES

[1] J. Lei, T. L. Berg, and M. Bansal, "Detecting moments and highlights in videos via natural language queries," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, no. 2021, pp. 11846–11858.

[2] K. Ma, X. Zang, Z. Feng, H. Fang, C. Ban, Y. Wei, Z. He, Y. Li, and H. Sun, "LLaViLo: Boosting video moment retrieval via adapter-based multimodal modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 2790–2795.

[3] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5277–5285.

[4] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 706–715.

[5] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu, "VTimeLLM: Empower LLM to grasp video moments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 14271–14280.

[6] A. Diwan, P. Peng, and R. Mooney, "Zero-shot video moment retrieval with off-the-shelf models," in *Proc. 2nd Workshop Trends Appl. Speech Language Technol. NLP (TL4NLP)*, Dec. 2022, pp. 24–30.

[7] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, "UniVTG: Towards unified video-language temporal grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 2782–2792.

[8] M. Zheng, L. Wang, X. He, C. Shen, and P. Luo, "Weakly supervised video moment localization with contrastive negative sample mining," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2022, vol. 36, no. 3, pp. 3517–3525.

[9] S. Kong, L. Li, B. Zhang, W. Wang, B. Jiang, C. Yan, and C. Xu, "Dynamic contrastive learning with pseudo-samples intervention for weakly supervised joint video MR and HD," in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, ON, Canada, Oct. 2023, pp. 538–546.

[10] G. Wang, X. Wu, Z. Liu, and J. Yan, "Prompt-based zero-shot video moment retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Lisbon, Portugal, Oct. 2022, pp. 413–421.

[11] J. Wattasseril, M. Faruqui, S. Pan, and A. G. Roy, "Zero-shot video moment retrieval using BLIP-based models," in Proc. Int. Symp. Vis. Comput. (ISVC), Lake Tahoe, NV, USA, Oct. 2023, pp. 160–171.

[12] Y. Xu, Y. Sun, B. Zhai, M. Li, W. Liang, Y. Li, and S. Du, "Zero-shot video moment retrieval via off-the-shelf multimodal large language models," 2025, *arXiv:2501.07972*.

[13] X. He, D. Yang, W. Feng, T.-J. Fu, A. Akula, V. Jampani, P. Narayana, S. Basu, W. Y. Wang, and X. Wang, "CPL: Counterfactual prompt learning for vision and language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 3407–3418.

[14] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 312–328.

[15] S. Yoon, J. W. Hong, S. Eom, H. S. Yoon, E. Yoon, D. Kim, J. Kim, C. Kim, and C. D. Yoo, "Counterfactual two-stage debiasing for video corpus moment retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[16] A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2127–2136.

[17] M. Narasimhan, A. Rohrbach, and T. Darrell, "CLIP-It! language-guided video summarization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 13988–14000.

[18] S. Yoon, J. Lee, K.-W. On, J. H. Kim, and B.-T. Zhang, "Selective query-guided debiasing for video corpus moment retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 185–200.

[19] C. Hur, J. H. Hong, D. H. Lee, D. Kang, S. Myeong, S. H. Park, and H. Park, "Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, Jun. 2025, pp. 24077–24086.

[20] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, Toronto, ON, Canada, Jul. 2024, pp. 12585–12602.

[21] W. Wu, H. Luo, B. Fang, J. Wang, and W. Ouyang, "Cap4 Video: What can auxiliary captions do for text-video retrieval?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 10704–10713.

[22] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8748–8763.

[23] S. Bai et al., "Qwen2.5-VL technical report," 2025, *arXiv:2502.13923*.

[24] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and R. Ganapathy, "The Llama 3 herd of models," 2024, *arXiv:2407.21783*.

[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164.

[26] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 2048–2057.

[27] Q. Li, B. Wang, and M. Melucci, "CNM: An interpretable complex-valued network for matching," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, Minneapolis, MN, USA, Jun. 2019, pp. 331–340.

[28] Y. Huang, L. Yang, and Y. Sato, "Weakly supervised temporal sentence grounding with uncertainty-guided self-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 18908–18918.

[29] J. Nam, D. Ahn, D. Kang, S. J. Ha, and J. Choi, "Zero-shot natural language video localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 1450–1459.

[30] J. Gao and C. Xu, "Learning video moment retrieval without a single annotated video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1646–1657, Mar. 2022.

[31] D. Luo, J. Huang, S. Gong, H. Jin, and Y. Liu, "Zero-shot video moment retrieval from frozen vision-language models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2024, pp. 5452–5461.

[32] Z. Li, Q. Xu, D. Zhang, H. Song, Y. Cai, Q. Qi, R. Zhou, J. Pan, Z. Li, V. Tu Vu, Z. Huang, and T. Wang, "GroundingGPT: Language enhanced multi-modal grounding model," 2024, *arXiv:2401.06071*.

[33] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "TimeChat: A time-sensitive multimodal large language model for long video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 14313–14323.

[34] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," 2023, *arXiv:2306.02858*.

[35] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie, "UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 20456–20465.

[36] P. Li, S. Wu, Y. Zhang, H. Wang, and J. Li, "MomentDiff: Generative video moment retrieval from random to real," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 65948–65966.
[37] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Advances in Neural Information Processing Systems (NeurIPS), 2023, pp. 34892–34916.
[38] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, Jul. 2023, pp. 19730–19742.
[39] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," Mistral AI, Paris, France, Tech. Rep. 1, 2023.



MINGYU JEON (Member, IEEE) received the B.S. degree in data science from Sejong University, Seoul, Republic of Korea, in 2024. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include video moment retrieval and various multimodal problems based on video.



MINUK MA received the B.S. and M.S. degrees from KAIST, Republic of Korea, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Ding Group, Department of Computer Science, The University of British Columbia. He was a Research Scientist and a Team Leader of the Data Centric AI Team, Oncology Research, Lunit Inc., Republic of Korea. His research interests include multi-agent AI systems, physics-informed deep learning, and protein–protein interactions.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Researcher Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual language reasoning, visual question answering, and various video-based problems.