Mitigating Modality and Language-Style Gaps for Zero-Shot Video Moment Retrieval

Jihyun Lee*, Cheol-Ho Cho*, Woojin Jun, Woojin Jeong, and Jae-Pil Heo**

Sungkyunkwan University, Suwon, Republic of Korea

Abstract. Zero-shot video moment retrieval (ZMR) aims to overcome the limitations of traditional approaches that require large-scale datasets annotated with text and its relevant temporal spans. Despite advances in pre-trained vision–language models (VLMs) and multimodal large language models (MLLMs), existing ZMR methods still heavily depend on query-to-video content similarity, making them vulnerable to modality and language-style gaps. These gaps lead to unreliable span proposals and unstable moment retrieval results. To address this issue, we propose Self-Similarity-based Moment proposal and Scoring (Self-SiMS) that instead exploits intrinsic relationships within videos, enabling robust span generation and scoring. By deriving self-similarity only from the video content, we circumvent the noisy and mismatched patterns of query–frame or query–caption similarities, thereby mitigating both modality and language-style gaps. Furthermore, we introduce a query-aware MLLM-based reasoning stage to further sharpen alignment between text and video. Extensive experiments demonstrate that Self-SiMS achieves the state-of-the-art performance across ZMR benchmarks.

1 Introduction

With the rapid expansion of multimedia content, video has become a primary medium for obtaining information. Consequently, retrieving specific moments from videos relevant to natural language queries has garnered significant attention across both academia and industry. A central task in this area is video moment retrieval (VMR) [2, 7, 15, 37], which aims to localize the most relevant temporal segment in a video based on a given text query. This task remains particularly challenging due to the need to pinpoint precise moments within videos that often contain multiple semantically similar scenes.

Previous studies have primarily focused on training VMR models using datasets annotated with temporal spans. However, such approaches require extensive manual labeling, which is both costly and labor-intensive. To overcome this limitation, recent research has explored zero-shot video moment retrieval (ZMR) [27,

* These authors contributed equally.
** Corresponding author

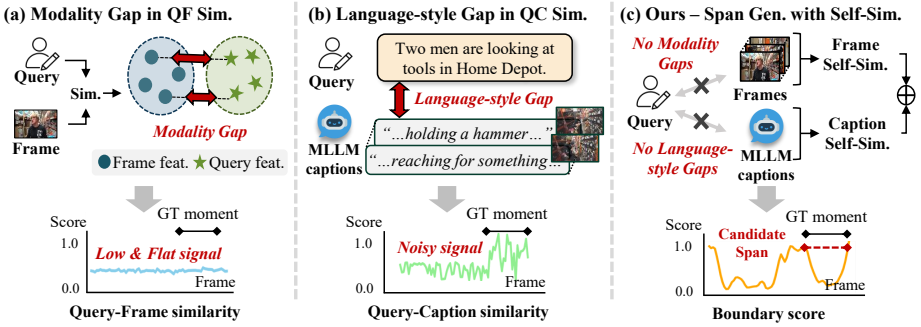

Fig. 1: Comparison of different similarity computation for candidate span generation. (a) Query–Frame similarity suffers from a modality gap between textual and visual features, producing low and flat similarity signals. (b) Query–Caption similarity alleviates modality issues but introduces a language-style gap, yielding noisy and inconsistent similarity patterns. (c) Our method leverages intra-video self-similarity, avoiding such modality and language-style gaps, and produces more stable boundary scores.

28,34,38,41], leveraging vision-language models (VLMs) [17,18] and multimodal large language models (MLLMs) [8, 32, 33] pretrained on large-scale data to perform retrieval without additional supervision. Although these methods have shown promising performance in ZMR, they still face a critical limitation stemming from the unreliable query-based similarity score used to generate and score candidate spans (i.e., moment proposals). Specifically, existing methods [38, 41] rely on the similarity between text query and video contents, represented by either visual features or MLLM-generated captions. However, these similarities are inherently affected by modality and language-style gaps, leading to the generation of unstable spans. Fig. 1 conceptually summarizes how each approach computes similarity scores and visualizes the score tendencies observed at the video instance level when applying each method.

As illustrated in Fig. 1(a), methods based on query–frame similarity [41] are affected by the modality gap between textual and visual representations. Even inside the ground-truth moment, the similarity can remain low and flat, making it difficult to distinguish the relevant temporal region from the surrounding context. Query–caption similarity [38] partially alleviates this issue by converting visual content into text, as shown in Fig. 1(b). However, it introduces a language-style gap between human-written queries and MLLM-generated captions. Even for visually similar frames, the captions may focus on different aspects of the same scene, such as objects, actions, or fine-grained sub-events. For example, given the query "Two men are looking at tools in Home Depot," captions may alternately emphasize "holding a hammer" or "reaching for something." These descriptions can be valid, but their inconsistent focus leads to noisy and unstable query–caption similarity trajectories.

To avoid relying on such query-dependent signals, we use query-independent intra-video self-similarity for span generation. As shown in Fig. 1(c), we compute self-similarity maps from both frame features and MLLM-generated caption features, and derive boundary scores by detecting changes in their temporal

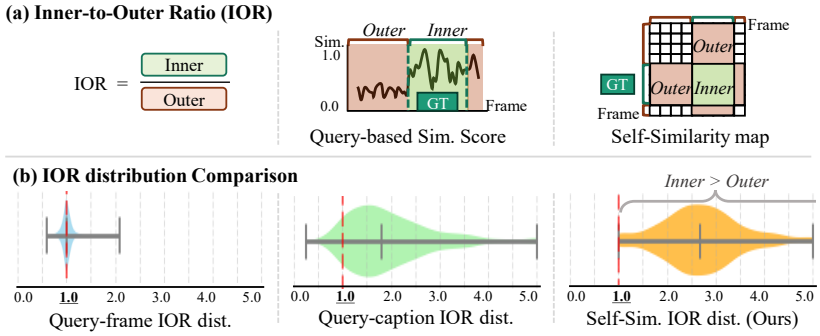

Fig. 2: Illustration and comparison of the Inner-to-Outer Ratio (IOR). (a) The IOR measures how much higher the similarity is within the ground-truth (Inner) region than in the surrounding (Outer) area. The middle and right examples illustrate how the Inner and Outer regions are obtained from query-based similarity scores and from the self-similarity map, respectively. (b) Comparison of IOR distributions on the QVHighlights validation set shows that our method achieves consistently higher IOR values than query–frame and query–caption similarities, indicating stronger discrimination between relevant and irrelevant segments.

similarity patterns. These boundary scores are then used to generate candidate spans. Since this process depends only on the internal structure of the video, rather than query–frame or query–caption matching, the resulting spans are less affected by modality and language-style gaps.

While these trends are discovered at the instance level, similar patterns emerge at the dataset level as well. To further examine these trends at the dataset level, we define the Inner-to-Outer Ratio (IOR) as a quantitative measure. Fig. 2(a) illustrates the definition of IOR, which measures how much higher the similarity is within the ground-truth (Inner) region than in the surrounding (Outer) area. An IOR above 1.0 indicates stronger similarity within the ground-truth span, whereas lower values imply weak discrimination, thereby measuring how well each method separates relevant from irrelevant segments. Fig. 2(b) reports the IOR distributions on the QVHighlights validation set as a representative example. The query–frame similarity exhibits IOR values concentrated around 1.0, corresponding to the flat similarity patterns observed earlier and indicating that the modality gap leads to weak discrimination between inner and outer regions. The query–caption similarity shows a broad and unstable distribution, reflecting the noisy similarity trajectories caused by inconsistent alignment between queries and generated captions. In contrast, our method yields IOR values consistently above 1.0, indicating that it effectively captures intrinsic video structure and produces discriminative similarity scores across the dataset.

Building on this idea, we propose Self-Similarity-based Moment Proposal and Scoring (Self-SiMS), which leverages intra-video self-similarity for both candidate generation and span scoring. For span generation, self-similarity provides query-independent temporal structure, avoiding unstable query-frame or query-caption signals. For span scoring, we first use the most reliable query-caption signals within each span and then refine the score with self-similarity to reflect

internal content consistency. This design reduces the effect of noisy query-caption alignment while preserving temporally coherent candidate spans.

Finally, to accurately select the final span from semantically similar span candidates, we introduce a query-aware MLLM-based re-ranking step. In this stage, representative frames from each span are directly evaluated against the query using an MLLM with a yes/no prompt. Instead of computing similarity after independently generating representations for the query and video, we leverage the MLLM's query-aware capabilities to further reduce both modality and language-style gaps, resulting in more accurate span scoring.

To sum up, our contributions are as follows:

- We first explicitly identify and address the modality and language-style gaps in zero-shot video moment retrieval (ZMR), where mismatches between text queries and video contents, whether visual features or MLLM-generated captions, can lead to inaccurate span scoring.
- To mitigate these gaps, we propose Self-Similarity-based Moment proposal and Scoring (Self-SiMS). To the best of our knowledge, our work is the first to leverage temporal self-similarity for generating candidate spans in a training-free ZMR, and to score these spans using intra-video relations.
- We further introduce a query-aware MLLM re-ranking step, where candidate spans are directly evaluated against the text query using an MLLM to mitigate the gaps above.
- Our method achieves the state-of-the-art performance on ZMR benchmarks, demonstrating the effectiveness of combining Self-SiMS with query-aware MLLM re-ranking in mitigating modality and language-style gaps and improving ZMR.

2 Related works

Video moment retrieval. Video moment retrieval (VMR) aims to retrieve relevant temporal segments from an untrimmed video given a natural language query. Early work treated this as a query-to-segment matching problem using joint video–text feature learning and alignment techniques [2,7,14]. With the introduction of large-scale datasets such as ActivityNet [3] and QVHighlights [15], subsequent methods explored attention mechanisms, cross-modal transformers, and contrastive learning for improved localization [15,18,24,31,37,39]. Recent approaches also leverage pre-trained vision–language models (VLMs) such as BLIP [18], BLIP-2 [17], and instruction-tuned video models [26,40], achieving strong results in moment retrieval and highlight detection. However, these advances still rely on large-scale span annotations, motivating research into weakly-supervised and zero-shot retrieval.

Zero-shot video moment retrieval. Traditional VMR approaches rely on densely annotated datasets that precisely align queries with video moments [15, 31], but such annotations are costly and labor-intensive, motivating research

into zero-shot video moment retrieval (ZMR) [27, 34]. Early zero-shot methods leveraged pretrained VLMs such as CLIP [28] to compute query–frame similarity without training data [27], yet they struggled to capture temporal relationships. Moreover, recent studies show that off-the-shelf multimodal large language models (MLLMs) can achieve competitive zero-shot performance without additional training [6,9,23,35]. Despite these advances, existing methods still rely on query–video content similarity using either visual features or MLLM-generated captions for candidate span generation and scoring, which makes them vulnerable to modality and language-style gaps and limits precise moment localization.

Modality and language-style gaps in text-visual alignment. Pretrained VLMs [17, 18, 28] trained on large-scale text–image pairs have advanced cross-modal understanding, yet a modality gap remains, causing inaccurate similarity matching. Prior works address this by introducing learnable gap parameters [36], enhancing semantic modeling [22], or generating captions with MLLMs to bridge text and visuals. However, MLLM-generated captions inherently create a language-style gap with human-written queries, which some studies mitigate by augmenting queries with complementary descriptions [20, 38]. Despite these prior efforts, ZMR still suffers from both gaps, often leading to unreliable span generation and scoring. Therefore, we propose a novel framework that leverages video-intrinsic self-relationships to mitigate them jointly.

Self-similarity for temporal boundary modeling. Self-similarity has been studied as a fundamental cue for capturing internal relational patterns in images and videos [29]. Building on this idea, recent works exploit temporal similarity to obtain pseudo temporal proposal during training. PSVL [27], a training-based MR method, constructs a frame-wise self-similarity matrix to discover pseudo event regions and pairs them with pseudo queries, which are then used to train a localization model. Similarly, another training-based video grounding framework [12] segments videos using a temporal similarity matrix and trains a grounding model by aligning the resulting pseudo proposals with pseudo language features in a pretrained vision-language space. In contrast, we alleviate modality and language-style gaps in a training-free manner: we derive query-agnostic candidate spans from intra-video self-similarity and perform span scoring, thereby reducing reliance on unstable query-dependent similarity scores.

3 Method

3.1 Overview

Given an untrimmed video $V = \{v_i\}_{i=1}^{L}$ with L sampled frames and a textual query Q, zero-shot video moment retrieval aims to localize the temporal span aligned with Q without requiring any training. Fig. 3 illustrates the overall pipeline of Self-SiMS. We first extract frame features F^f and frame-wise caption features F^c, and use them to build two temporal self-similarity maps, M^f and M^c. These maps are merged into a unified self-similarity map M, from which

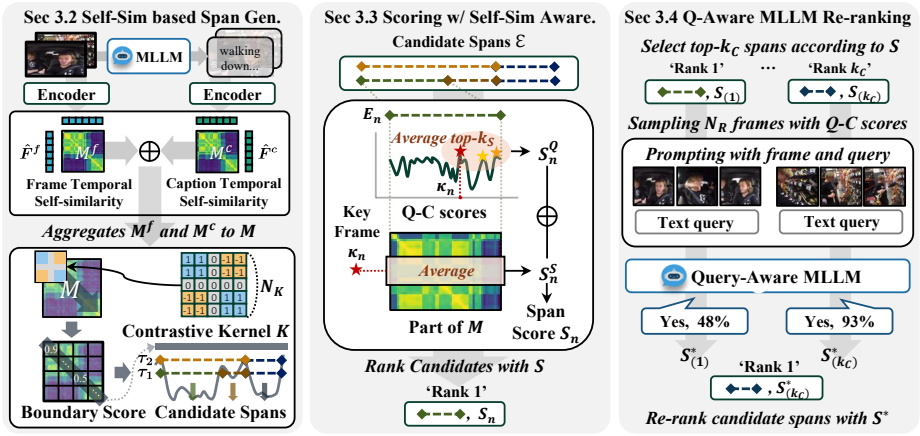

Fig. 3: The overview of our Self-Similarity based Moment proposal and Scoring (Self-SiMS). First, features from the frames and captions are extracted to construct two distinct self-similarity maps, M^f and M^c, which are then aggregated into a unified map, M. Each frame is then evaluated with a boundary score computed using a contrastive kernel, and candidate spans are generated in temporal order (Sec. 3.2). Next, the spans are ranked based on query-matching span score S^Q, which is computed by averaging the top-k_S query-caption scores. In addition, the self-matching span score S^S is derived from the unified Self-similarity map M and then combined with the former score (Sec. 3.3). Finally, the MLLM-based reranking is applied to the top-ranked spans by asking whether each query-frame pair is relevant, and the probability of the 'Yes' answer is used as the score (Sec. 3.4).

we compute boundary scores with a contrastive kernel and generate candidate spans. Each candidate is then scored by combining the Query-Matching Span Score S^Q, computed from the top-k_S query–caption similarities within the span, and the Self-Matching Span Score S^S, which measures internal span consistency using M. The combined score S ranks the candidates, and the top-k_C spans are further re-ranked by an MLLM through a query-aware Yes/No verification.

Span Boundary Scores. Inspired by the event boundary detection method proposed in [11], we calculate self-similarity to construct the Temporal Self-similarity Matrix (TSM), $M \in \mathbb{R}^{L \times L}$. First, we apply ℓ_2-normalization to the extracted features, F^f for frames and F^c for captions, resulting in $\hat{F}^f$ and $\hat{F}^c$, respectively. Then, the TSMs for the frames and captions are computed as $M^f = \hat{F}^f(\hat{F}^f)^\top$ and $M^c = \hat{F}^c(\hat{F}^c)^\top$. To integrate both modalities, we construct the unified TSM M by equally weighting M^f and M^c, i.e., $M = \frac{1}{2}(M^f + M^c)$.

To evaluate whether a frame is likely to be a boundary, we examine a local patch of the self-similarity map around that frame. Intuitively, a good boundary separates two temporally coherent regions. Frames before the boundary should be similar to each other, frames after the boundary should also be similar to each other, but frames across the boundary should be less similar. The contrastive kernel captures this pattern by rewarding within-side similarities and penalizing cross-side similarities. Specifically, the kernel $K \in \mathbb{R}^{N_K \times N_K}$ has positive weights

in the top-left and bottom-right quadrants, negative weights in the top-right and bottom-left quadrants, and zeros in the central row and column. The boundary score of the i-th frame is computed as

$$b_i = P_i \odot K, \tag{1}$$

where P_i is an $N_K \times N_K$ patch of M centered at the diagonal position (i, i), and $\odot$ denotes element-wise multiplication. The TSM is zero-padded so that this operation can be applied uniformly to every frame.

Candidate Spans. To generate candidate spans from the boundary scores, we define a set of instance-wise dynamic thresholds $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_{N_\mathcal{T}}\}$, which adapt to the intra-video context variability measured from the frame features F^f. Based on this variability, videos with more diverse contexts are assigned lower thresholds for finer temporal partitioning, whereas more homogeneous videos receive higher thresholds for coarser segmentation. This variability-aware design removes the need for manual threshold tuning and generalizes well across datasets. Detailed formulations are provided in the Appendix. For each $\tau_j \in \mathcal{T}$, we construct boundary set $\mathcal{B}_j$ containing the indices of frames that are both above the threshold and correspond to local maxima of the boundary scores:

$$\mathcal{B}_j = \{\, i \mid b_i \geq \tau_j,\ b_i \geq b_{i-1},\ b_i \geq b_{i+1} \,\} \cup \{1, L\}. \tag{2}$$

Here, the two extreme frames, 1 and L, are included to ensure that the candidate spans cover the entire duration of the video. By the boundary set, we obtain a candidate span E_n formally defined as:

$$E_n = \{i \mid e_n^s \leq i \leq e_n^e\}, \quad 1 \leq e_n^s < e_n^e \leq L, \tag{3}$$

where each pair (e_n^s, e_n^e) is formed by two successive frame indices in the temporal order within the same boundary set $\mathcal{B}_j$, thereby ensuring that the resulting different spans do not overlap. Finally, the set of all candidate spans across thresholds is formally denoted by

$$\mathcal{E} = \{E_1, E_2, \ldots, E_{N_s}\}, \tag{4}$$

where N_s denotes the total number of candidate spans. Note that each E_n in this set refers to the corresponding candidate span and can be referred to as the "n-th span" in subsequent descriptions.

3.2 Span Scoring with Self-Similarity Awareness

To identify the relevant candidate span, each span should have its relevance score with the query. Previous approach [38] computes this score by averaging the frame-level similarities between the query and captions. However, such mean-based scoring is unreliable, as query–caption similarity often exhibits noisy and inconsistent distributions due to the language-style gap.

To mitigate this limitation, we use the top-k_S frame-level similarities within the span for scoring, which are robust to the corruption induced by the language-style gap. In addition, to enhance the reliability of scoring, we propose the Self-Matching Span Score, derived from TSM. Finally, each candidate span is scored by combining the Query-Matching Span Score and the Self-Matching Span Score.

Query–Matching Span Score. Given a textual query Q, we extract the query feature F^q using the same encoder employed for the caption features F^c. Then, the cosine similarity between the query feature and caption features is calculated to produce the frame-level scores $S^f \in \mathbb{R}^L$ as follows:

$$S_i^f = \frac{F^q \cdot F_i^c}{\|F^q\| \, \|F_i^c\|}, \tag{5}$$

where F_i^c and S_i^f denote the caption feature and frame-level score of the i-th frame, respectively. The Query-Matching Span Score (QMS), denoted as S^Q, is defined as the mean of the top-k_S frame scores within each span. Specifically, for the n-th span E_n, containing the indices from the start to the end frames e_n^{s} and e_n^{e}, the score is computed as follows:

$$S_n^Q = \frac{1}{k_S} \sum_{i=e_n^{\mathrm{s}}}^{e_n^{\mathrm{e}}} \mathbf{1}\left[i \in I_{k_S}^{E_n}\right] S_i^f, \tag{6}$$

where $I_{k_S}^{E_n}$ is a set of indices with the top-k_S highest frame-level scores defined in Eq. (5) within E_n. This approach prioritizes the most relevant frames over the average, helping to mitigate noise in the query–caption similarity by using the most reliable scores. As a result, the span score more reliably reflects each span's relevance to the query.

Self-Matching Span Score. However, relying solely on a subset of sampled frames inherently limits the ability to capture alignment between the query and the full temporal context of a span. To address this issue, we leverage the previously computed TSM used in span generation to evaluate contextual consistency within the span while reducing the influence of the language-style gap. The Self-Matching Span Score (SMS), denoted as S^S, is defined as the mean similarity between the key frame, which is the frame with the highest frame-level score in the span, and all other frames within that span. Let κ_n denote the index of the key frame in the span E_n, that is,

$$\kappa_n = \arg\max_{i \in E_n} S_i^f. \tag{7}$$

Then, SMS for the n-th candidate span is calculated as

$$S_n^S = \frac{1}{|E_n|} \sum_{j \in E_n} M_{\kappa_n, j}, \tag{8}$$

where $M_{i,j}$ denotes the element of the TSM M at the position (i, j). As a result, candidate spans that are longer yet include inconsistent contexts are suppressed through reduced span scores. Finally, the total score S_n of the candidate span is weighted sum of S_n^Q and S_n^S as follows:

$$S_n = (1 - \alpha)\, S_n^Q + \alpha\, S_n^S, \quad \alpha \in [0, 1], \tag{9}$$

where α is a hyperparameter for the span score weights.

3.3 Query-Aware Gap Reduction via MLLM Re-Ranking

Recent remarkable advances in multimodal large language models (MLLMs) have demonstrated exceptionally strong cross-modal reasoning ability, making them effective for re-ranking tasks [5,10,19,30]. Building on this, we further reduce the language-style gap between queries and MLLM-generated captions by refining candidate spans through MLLM-based re-ranking.

Concretely, we re-rank the top-k_C candidate spans from the initial scoring stage, ranked according to their span score S. Let $S_{(m)}$ and $E_{(m)}$ denote the m-th highest span score and its corresponding candidate span, where $1 \leq m \leq k_C$. Within each selected span, we sample N_R representative frames according to their query–caption similarity score. Further details of the N_R frame sampling process are provided in the Appendix. Then, each representative frame is evaluated by the MLLM using a *Yes/No* prompt, and the resulting logits are converted into probabilities via softmax. The resulting *Yes* probability is used as S_i^r, denoting the frame-level re-ranking score for the i-th frame. By averaging the top-k_R highest values from the set of frame-level re-ranking scores S_i^r, we compute the span-level re-ranking score $S_{(m)}^R$. Finally, $S_{(m)}^*$, the final score of $E_{(m)}$, is defined as a weighted combination of the initial span score $S_{(m)}$ and the MLLM-based re-ranking signal $S_{(m)}^R$:

$$S_{(m)}^* = (1 - \beta)\, S_{(m)} + \beta\, S_{(m)}^R, \quad \beta \in [0, 1], \tag{10}$$

where β is a weighting parameter. This re-ranking process leverages the MLLM's cross-modal reasoning to directly verify semantic alignment between the query and candidate captions, rather than relying solely on similarity scores, enabling query-aware reasoning and more reliable selection of the most relevant spans.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on four widely used benchmarks. QVHighlights [15] includes user-generated videos with fine-grained highlight annotations. Charades-STA [7] contains indoor daily activities paired with temporal sentence

Method	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
Moment-DETR [15]	FS	52.9	33.0	54.8	30.7	54.2	33.4	55.4	31.1
UMT [21]	FS	56.4	40.8	53.1	35.4	-	-	-	-
CPL [43]	WS	30.8	10.8	22.8	-	-	-	-	-
CPI [13]	WS	32.3	11.8	23.7	-	-	-	-	-
TFVTG‡ [41]	ZS	21.6	8.0	20.1	7.9	21.0	7.4	19.2	7.3
Moment-GPT [38]	ZS	58.3	37.7	55.1	35.0	58.9	38.6	55.7	35.9
Self-SiMS (Ours)	ZS	**59.7**	**42.2**	**59.2**	**38.3**	**61.0**	**43.2**	**60.5**	**39.3**

Table 1: Comparison on QVHighlights test and validation sets. FS: Fully Supervised, WS: Weakly Supervised, ZS: Zero-Shot. ‡: Reproduced with the same LLM and FPS with ours.

descriptions. ActivityNet-Captions [14] covers diverse human activities with temporally aligned captions. Additionally, we incorporate TVR [16], which features long-form videos and complex natural language queries, to further validate the generalizability of our approach across varied video–language domains.

Metrics. Following prior zero-shot video moment retrieval (ZMR) works, we evaluate our method using standard metrics tailored to each dataset. For QVHighlights, we report Recall@1 at IoU thresholds of 0.5 and 0.7, along with mAP@0.5 and the average mAP across multiple thresholds. These metrics jointly reflect the retrieval accuracy and the quality of segment ranking. For Charades-STA, ActivityNet-Captions, and TVR, we adopt R1@0.3/0.5/0.7 and mean IoU (mIoU), which evaluate both temporal localization precision and the alignment between predicted and ground-truth segments.

Implementation Details. We set the frame rates of QVHighlights, Charades-STA, ActivityNet-Captions, and TVR to 0.5, 1, 1, and 1 fps, respectively. Following prior work [38], we use 0.5 fps for QVHighlights, and we set the remaining datasets to 1 fps for consistency. We employ LLaMA-3.2-11B-Vision-Instruct [8] for both video frame captioning and the final re-ranking stage. For scoring, the number of top scores is fixed at 3 for both initial span scoring (k_S) and re-ranking (k_R). We re-rank the top-k_C candidate spans with $k_C = 5$. In the re-ranking stage, N_R frames are sampled from each span, set to half of the span length with lower and upper bounds of 10 and 30, respectively. The weighting parameters for the final span score, α and β, are set to 0.1 and 0.5, respectively. All experiments are conducted on an NVIDIA RTX A6000 GPU.

4.2 Comparison With the State-of-the-Arts

Tabs. 1,2, and 3 report the performance of various ZMR methods across benchmark datasets. In Tab. 1, Self-SiMS outperforms existing ZMR baselines across all evaluation metrics on the QVHighlights. Note that TFVTG‡ was reproduced using the same large language model (LLM) employed in our framework, since TFVTG does not officially provide LLM rephrasing output for QVHighlights.

Tab. 2 presents the performance comparison on Charades-STA and ActivityNet-Captions. To ensure fairness, we reproduced TFVTG† at 1 fps, as the original

Method	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
VTimeLLM [9]	FS	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
Moment-DETR [15]	FS	62.1	48.2	25.3	42.3	52.6	32.5	15.3	37.8
CNM [42]	WS	50.0	36.2	14.2	34.2	51.3	30.3	11.4	33.9
CPL [43]	WS	56.0	38.1	20.3	37.8	52.4	30.9	12.0	32.6
TFVTG† [41]	ZS	<u>64.7</u>	29.8	9.8	38.2	49.9	26.8	12.6	<u>34.2</u>
TFVTG‡ [41]	ZS	**64.8**	30.0	10.1	<u>38.4</u>	49.5	26.5	12.3	34.0
Moment-GPT [38]	ZS	58.2	<u>38.4</u>	**21.6**	36.5	48.1	**31.1**	**14.9**	30.8
Self-SiMS (Ours)	ZS	62.7	**39.7**	<u>21.0</u>	**41.9**	**49.9**	<u>28.2</u>	<u>13.8</u>	**34.7**

Table 2: Comparative evaluation on Charades-STA and ActivityNet-Captions. ZS: Zero-Shot. †: Reproduced results with the same FPS with ours. ‡: Reproduced with the same LLM and FPS with ours.

Method	TVR					
	s	f	R1@0.3	R1@0.5	R1@0.7	mIoU
TFVTG‡ [41]	40	1.0	13.1	6.0	3.0	15.1
	50	0.5	14.4	6.9	2.7	15.4
	20	0.5	21.7	9.6	4.0	18.6
Ours	-	-	**26.6**	**13.5**	**6.1**	**20.2**

Table 3: Comparison on TVR validation dataset. For the baseline methods, s (stride) and f (max stride factor) are dataset-specific hyperparameters as reported in TFVTG [41].

results were reported at 3 fps. TFVTG achieves relatively high R@0.3 and mIoU but exhibits clear drops at higher IoU thresholds (R1@0.5 and R1@0.7), as its query–visual similarity often results in overly long spans. Conversely, Moment-GPT performs better at higher thresholds but worse at R@0.3 and mIoU, indicating that its query–caption similarity produces precise yet unstable short matches. In contrast, our Self-SiMS overcomes these issues by producing more consistent spans aligned with the ground-truth, achieving the highest mIoU and more reliable temporal localization while maintaining competitive recall.

Tab. 3 summarizes the evaluation on the TVR dataset. As the official implementation of Moment-GPT is unavailable, we omit its results for fairness. For TFVTG, we report multiple configurations using the hyperparameters s (stride) and f (max stride factor) as specified in the original paper, ensuring a fair comparison. Our Self-SiMS consistently surpasses the baseline across all parameter settings, marking a substantial improvement. Moreover, while the baseline requires dataset-specific hyperparameter tuning, Self-SiMS delivers strong performance with a single configuration, highlighting its robustness.

4.3 Ablation Studies

Length of Candidate Spans Analysis.

Fig. 4 illustrates the candidate span length distributions on Charades-STA (left) and ActivityNet-Captions (right). The ground-truth (GT) distribution is indicated by the red dashed line. This analysis helps explain the performance

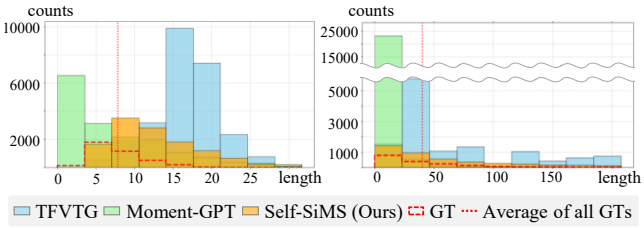

Fig. 4: Histogram of all span lengths on Charades-STA (left) and ActivityNet-Captions (right).

Method	Oracle-mIoU (%)	Avg. # Span Cand.
TFVTG	38.16	8.38
Moment-GPT	65.01	7.62
Self-SiMS (Ours)	**71.13**	**6.46**

Table 4: Comparison of Oracle mIoU on the QVHighlights validation set. Avg. # Span Cand. denotes the average number of span candidates across on dataset.

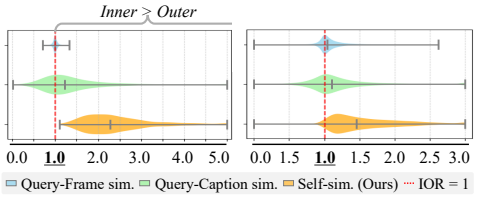

Fig. 5: Inner-to-Outer Ratio (IOR) distributions on Charades (left) and ActivityNet-Captions (right).

differences observed on these two datasets, particularly the recall tendencies in Tab. 2. TFVTG, affected by the modality gap, produces uniform similarity scores with weak temporal distinction. Consequently, it often produces excessively long spans, achieving relatively high R@0.3 but exhibiting clear degradation under stricter IoU thresholds. Conversely, Moment-GPT suffers from a language-style gap, leading to noisy and unstable similarity score distribution. These sharp, unstable peaks often yield overly short spans that perform better under tight criteria such as R@0.7 yet fail to capture full GT intervals. In contrast, our Self-SiMS generates balanced spans that align more closely with the GT distribution by leveraging self-similarity free from such modality and style gaps, resulting in higher mIoU and more consistent temporal localization performance.

Effectiveness of Span Generator. To validate the effectiveness of our span generator, we evaluate Oracle mIoU on the QVHighlights, comparing against TFVTG [41] and Moment-GPT [38]. Oracle mIoU is defined as the mean IoU obtained by selecting the candidate span with the maximum overlap with the ground truth, reflecting the quality of the span candidates. As shown in Tab. 4, our method achieves an Oracle mIoU of 71.13, outperforming both baselines. Moreover, although this performance generally increases with a larger number of candidate spans, our method achieves the best results despite requiring the lowest average number of candidates (6.46). This improvement demonstrates that our method generates more reliable spans by capturing the intrinsic structure of the video while avoiding modality and language-style gaps.

Inner-to-Outer Ratio Distribution. Fig. 5 reports IOR distributions on Charades-STA and ActivityNet-Captions, where higher IOR values indicate stronger separation between the ground-truth span and surrounding background. Self-

Method	R1@0.5	R1@0.7	mAP@0.5	mAP@avg
Mean scoring	55.7	40.9	56.7	37.1
QMS	59.5	42.1	59.4	38.6
+ SMS	60.3	42.9	59.6	38.7
+ Re-ranking	**61.0**	**43.2**	**60.5**	**39.3**

Table 5: Scoring component analysis of our scoring methods. QMS denotes the Query-Matching Score, and SMS denotes the Self-Matching Span Score.

Model	Component	MLLM / LLM	#Params	Runtime
TFVTG	Rephrasing	GPT4-Turbo [1]	$\geq$ 100B (est.)	–
Moment-GPT	Rephrasing	LLaMA3 [8]	8B	6.3s
	Captioning	MiniGPT-v2 [4]	7B	249.4s
	Reranking	V-ChatGPT [25]	7B	7.4s
Ours	Captioning	LLaMA3.2-V [8]	11B	140.7s
	Reranking	LLaMA3.2-V [8]	11B	13.0s
	Captioning	MiniGPT-v2 [4]	7B	249.4s
	Reranking	MiniGPT-v2 [4]	7B	9.7s

Table 6: Comparison of MLLM/LLM usage, model scale, and per-video runtime across TFVTG, Moment-GPT, and our method.

SiMS shows higher and more stable IOR values, demonstrating that its self-similarity-based boundary scores capture temporal segments more reliably and lead to robust localization.

Scoring Component Analysis. We conduct an ablation study on the QVHighlights validation set to analyze the contribution of each component in our proposed scoring and re-ranking pipeline, with the results presented in Tab. 5. The baseline adopts naive mean scoring, which averages all frame-level scores within a span, as in prior methods. Replacing mean scoring with the Query-Matching Span Score (QMS) yields improvements across all metrics, supporting our hypothesis that employing the most relevant signals is more effective than relying on the average of noisy query-caption scores. Adding the Self-Matching Span Score (SMS) on QMS further increases performance, particularly for recall-based metrics such as R1@0.5 and R1@0.7, indicating that incorporating the entire span context through self-similarity enables more precise localization of ground-truth moments. Finally, incorporating MLLM re-ranking on top of QMS with SMS yields additional gains across all metrics by leveraging the MLLM's ability to understand cross-modal information. This demonstrates that each component of the pipeline contributes effectively to the overall performance.

Complexity comparison. We compare the per-video runtime of TFVTG [41], Moment-GPT [38], and our method using 100 randomly sampled videos from Charades-STA [7]. Since non-MLLM/LLM components, such as our scoring stage, are lightweight (0.194s per video), we focus on MLLM/LLM-dependent stages and use a frame batch size of 1 for both captioning and reranking. Tab. 6 summarizes the model configurations and runtimes. TFVTG uses GPT-4 Turbo [1] for query rephrasing. Its runtime cannot be directly measured because it is accessed through a commercial API, and its model size is estimated to exceed 100B parameters. Moment-GPT employs three separate models for query rephrasing, captioning, and reranking, resulting in a total footprint of 22B parameters. In contrast, our default setting uses a single LLaMA-3.2-Vision [8] model family for both captioning and reranking, reducing the model footprint to 11B parameters. Furthermore, reranking is computationally efficient because it evaluates only a small number of sampled representative frames. To ensure a fair comparison under the same MLLM, we additionally replace LLaMA-3.2-Vision with MiniGPT-v2, which is also adopted by Moment-GPT. Under this

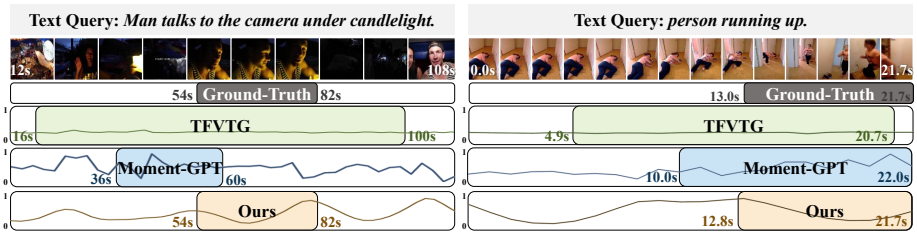

Fig. 6: Qualitative results on QVHighlights (left) and Charades-STA (right).

setting, the captioning cost is identical to that of Moment-GPT, while the reranking stage is slightly faster than when using LLaMA-3.2-Vision. Detailed performance results obtained with MiniGPT-v2 are provided in the supplementary MLLM ablation study. Notably, our method achieves comparable or superior performance while using the same MLLM backbone.

4.4 Qualitative results

We provide a qualitative comparison in Fig. 6, illustrating span localization results, TFVTG [41], Moment-GPT [38], and our method, along with the corresponding span candidate score distributions. TFVTG, which relies on query–visual frame similarity, generates overly long and inaccurate spans due to the modality gap that hinders fine-grained semantic alignment. Moment-GPT, which instead depends on query–caption similarity, suffers from a language-style gap that produces inconsistent similarity maps and hampers accurate localization. In contrast, Self-SiMS leverages boundary scores derived from the self-similarity, enabling robust localization of the correct spans.

5 Conclusion

In this paper, we first define the modality and language-style gaps in zero-shot video moment retrieval (ZMR), and quantify their impact using the Inner-to-Outer Ratio (IOR), which measures how well a similarity score separates the ground-truth span from the surrounding background. To mitigate these gaps, we propose Self-Similarity-based Moment Proposal and Scoring (Self-SiMS), which leverages intra-video self-similarity for both span generation and scoring. Experiments and ablations show that Self-SiMS mitigates errors from unreliable query-dependent similarity and achieves robust performance, demonstrating self-similarity as an effective cue for training-free ZMR.

Acknowledgements

This work was supported in part by MSIT/IITP (No. RS-2022-II220680, RS-2020-II201821, RS-2019-II190421, RS-2024-00459618, RS-2024-00360227, RS-2024-00437633, RS-2024-00437102, RS-2025-25442569), MSIT/NRF (No. RS-2024-00357729), KNPA/KIPoT (No. RS-2025-25393280), and SEMES-SKKU collaboration funded by SEMES.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Barrett, D.P., Barbu, A., Siddharth, N., Siskind, J.M.: Saying what you're looking for: Linguistics meets video search. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2069–2081 (2015)
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
4. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
5. Chen, Z., Xu, C., Qi, Y., Guo, J.: Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. arXiv preprint arXiv:2407.21439 (2024)
6. Diwan, A., Peng, P., Mooney, R.: Zero-shot video moment retrieval with off-the-shelf models. In: Transfer Learning for Natural Language Processing Workshop. pp. 10–21. PMLR (2023)
7. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
10. Jin, C., Peng, H., Zhao, S., Wang, Z., Xu, W., Han, L., Zhao, J., Zhong, K., Rajasekaran, S., Metaxas, D.N.: Apeer: Automatic prompt engineering enhances large language model reranking. In: Companion Proceedings of the ACM on Web Conference 2025. pp. 2494–2502 (2025)
11. Kang, H., Kim, J., Kim, T., Kim, S.J.: Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20073–20082 (2022)
12. Kim, D., Park, J., Lee, J., Park, S., Sohn, K.: Language-free training for zero-shot video grounding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2539–2548 (2023)
13. Kong, S., Li, L., Zhang, B., Wang, W., Jiang, B., Yan, C., Xu, C.: Dynamic contrastive learning with pseudo-samples intervention for weakly supervised joint video mr and hd. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 538–546 (2023)
14. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
15. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)

16. Lei, J., Yu, L., Berg, T.L., Bansal, M.: Tvr: A large-scale dataset for video-subtitle moment retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 447–463. Springer (2020)
17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
19. Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms. arXiv preprint arXiv:2411.02571 (2024)
20. Liu, D., Qu, X., Dong, J., Nan, G., Zhou, P., Xu, Z., Chen, L., Yan, H., Cheng, Y.: Filling the information gap between video and query for language-driven moment retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4190–4199 (2023)
21. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3042–3051 (2022)
22. Liu, Z., Li, J., Xie, H., Li, P., Ge, J., Liu, S.A., Jin, G.: Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 3855–3863 (2024)
23. Luo, D., Huang, J., Gong, S., Jin, H., Liu, Y.: Zero-shot video moment retrieval from frozen vision-language models. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5464–5473 (2024)
24. Ma, K., Zang, X., Feng, Z., Fang, H., Ban, C., Wei, Y., He, Z., Li, Y., Sun, H.: Llavilo: Boosting video moment retrieval via adapter-based multimodal modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2798–2803 (2023)
25. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
26. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
27. Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1470–1479 (2021)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
30. Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., Ren, Z.: Is chatgpt good at search? investigating large language models as re-ranking agents. arXiv preprint arXiv:2304.09542 (2023)

31. Sun, Y., Xu, Y., Xie, Z., Shu, Y., Du, S.: Gptsee: Enhancing moment retrieval and highlight detection via description-based similarity features. IEEE Signal Processing Letters **31**, 521–525 (2023)
32. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
33. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
34. Wang, G., Wu, X., Liu, Z., Yan, J.: Prompt-based zero-shot video moment retrieval. In: Proceedings of the 30th ACM international conference on multimedia. pp. 413–421 (2022)
35. Wattasseril, J.I., Shekhar, S., Döllner, J., Trapp, M.: Zero-shot video moment retrieval using blip-based models. In: International Symposium on Visual Computing. pp. 160–171. Springer (2023)
36. Xiao, J., Song, Z., Hu, J., Cheng, H., Hu, Z., Li, J., Hong, R.: Contrastive alignment with semantic gap-aware corrections in text-video retrieval. arXiv preprint arXiv:2505.12499 (2025)
37. Xu, Y., Sun, Y., Zhai, B., Jia, Y., Du, S.: Mh-detr: Video moment and highlight detection with cross-modal transformer. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
38. Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., Du, S.: Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8978–8986 (2025)
39. Xu, Y., Sun, Y., Zhai, B., Xie, Z., Jia, Y., Du, S.: Multi-modal fusion and query refinement network for video moment retrieval and highlight detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
40. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
41. Zheng, M., Cai, X., Chen, Q., Peng, Y., Liu, Y.: Training-free video temporal grounding using large-scale pre-trained models. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)
42. Zheng, M., Huang, Y., Chen, Q., Liu, Y.: Weakly supervised video moment localization with contrastive negative sample mining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3517–3525 (2022)
43. Zheng, M., Huang, Y., Chen, Q., Peng, Y., Liu, Y.: Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15555–15564 (2022)