Pattern Recognition 166 (2025) 111670

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

MLLM as video narrator: Mitigating modality imbalance in video moment retrieval

Weitong Cai [a], Jiabo Huang [a], Shaogang Gong [a], Hailin Jin [b], Yang Liu [c,*]

[a] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK
[b] Adobe Research, San Jose, CA, 95110, USA
[c] Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871, China

ARTICLE INFO

Keywords:
Video moment retrieval
Multi-modal large language model
Video search
Multi-modal learning

ABSTRACT

Video Moment Retrieval (VMR) aims to localize a specific temporal segment within an untrimmed long video given a natural language query. Existing methods often suffer from inadequate training annotations, *i.e.*, the sentence typically matches with a fraction of the prominent video content in the foreground with limited wording diversity. This intrinsic modality imbalance leaves a considerable portion of visual information remaining unaligned with text. It confines the cross-modal alignment knowledge within the scope of a limited text corpus, thereby leading to sub-optimal visual-textual modeling and poor generalizability. By leveraging the visual-textual understanding capability of multi-modal large language models (MLLM), in this work, we propose a novel MLLM-driven framework Text-Enhanced Alignment (TEA), to address the modality imbalance problem by enhancing the correlated visual-textual knowledge. TEA takes an MLLM as a video narrator to generate plausible textual descriptions of the video, thereby mitigating the modality imbalance and boosting the temporal localization. To effectively maintain temporal sensibility for localization, we design to get text narratives for each certain video timestamp and construct a structured text paragraph with time information, which is temporally aligned with the visual content. Then we perform cross-modal feature merging between the temporal-aware narratives and corresponding video temporal features to produce semantic-enhanced video representation sequences for query localization. Subsequently, we introduce a uni-modal narrative-query matching mechanism, which encourages the model to extract complementary information from contextual cohesive descriptions for improved retrieval. Extensive experiments on two benchmarks show the effectiveness and generalizability of our proposed method.

1. Introduction

Video Moment Retrieval (VMR) targets to identify moments of interest within untrimmed videos by predicting their temporal boundaries based on natural language query sentences describing specific activities [1–3]. This task poses a significant challenge in video understanding [4,5], requiring accurate comprehension of both visual and textual modalities and their precise alignment.

However, the cross-modal alignment knowledge provided by video-query training samples in existing datasets [6,7] is not always adequate in both *semantics completeness* and *diversity* to facilitate precise and generalizable moment-text correlation learning. Specifically, firstly, in terms of *semantic completeness*, as shown in Fig. 1(a), queries describing the user's moments of interest often only capture a fraction of the video segment content, focusing primarily on partial foreground elements rather than encompassing all the information within that moment (*e.g.*, failing to mention details such as 'mirror' and 'pile of clothes'). Additionally, despite the query being partially associated with its corresponding moments, the extensive visual content outside the moment of interest within the same video lacks textual descriptions, thus unable to contribute to bridging cross-modal understanding (*e.g.*, actions like 'lying on a bed' and 'standing in front of a bed'). Secondly, in terms of *diversity*, queries within a particular domain or dataset often tend to employ fixed words and phrases to convey similar or identical semantics [8,9]. This results in a fragile visual-textual understanding when confronted with various expressions. In summary, when contrasting with the richness of the video modality, the absence of a correlated text corpus, both in terms of semantic completeness and diversity, results in sub-optimal cross-modal learning for the VMR model. This intrinsic modality imbalance problem restricts the available

* Corresponding author.
E-mail addresses: weitong.cai@qmul.ac.uk (W. Cai), jiabo.huang@qmul.ac.uk (J. Huang), s.gong@qmul.ac.uk (S. Gong), hljin@adobe.com (H. Jin), yangliu@pku.edu.cn (Y. Liu).

https://doi.org/10.1016/j.patcog.2025.111670
Received 12 November 2024; Received in revised form 4 March 2025; Accepted 31 March 2025
Available online 11 April 2025

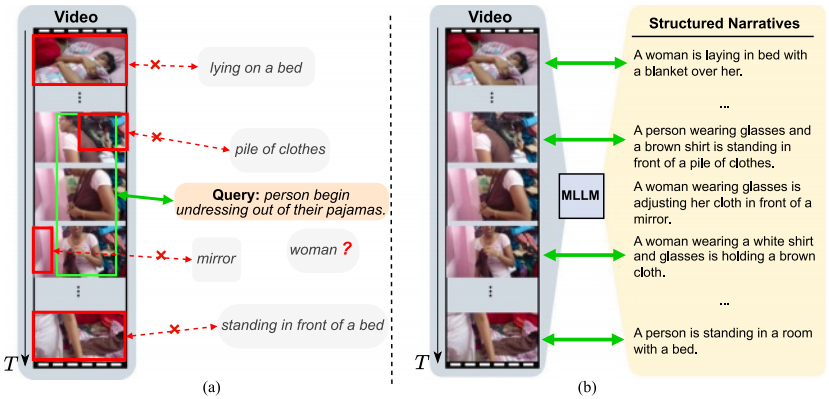

Fig. 1. An illustration of the intrinsic modality imbalance problem in video-query samples. (a) Query in existing datasets solely captures a fraction of the prominent video content *(semantic completeness)* in the foreground with the limited wording *diversity*, leaving a significant amount of visual information unaligned with text. (b) We leverage an MLLM as a video narrator to generate structured narratives temporally aligned with the corresponding video, to enhance the cross-modal understanding with the rich text corpus to facilitate more accurate and generalized predictions.

multi-modal alignment information to a limited text corpus, posing a risk of diminished generalization ability across various video-text correlations and distributions.

Existing attempts aiming at enhancing existing textual queries to augment correlated cross-modal information are broadly categorized into two strategies. One entails adjusting the syntax and/or wording of the ground-truth query to generate additional pairs and seeks to improve visual-textual alignment through contrastive learning [10–12]. However, it relies on strong assumptions about the known and consistent characteristics of syntax or wording and tailors rules to such specific traits, which may falter when confronted with different distributions. Another strategy involves leveraging manually aligned context from temporally neighboring ground-truth sentences within the same video to provide supplemental information [3,13,14]. However, depending on neighboring labeled sentences is frequently impractical in real-world scenarios with limited cross-modal annotations, thus constraining its scalability. It is worth noting that both strategies still rely on the limited associated visual-textual information within datasets through different permutations or combinations, which perpetuates the modality imbalance problem inherent in the original annotations.

Recent advancements in Multi-modal Large Language Models (MLLM) [15,16] have showcased the efficacy of text prompts in facilitating textual and visual comprehension and reasoning. Nevertheless, currently, many MLLMs are good at capturing the global visual semantics while hard to associate moments with accurate timestamps directly, due to the compression of visual inputs and limited grounding annotations [17,18]. One potential approach to leverage cross-modal knowledge is to directly generate corresponding queries from videos to create new training pairs. However, this task is non-trivial due to the ambiguity of moment boundaries [9]. Determining the endpoints of activities without human supervision introduces significant uncertainty and amplifies additional noise.

In this work, we propose a novel MLLM-driven VMR framework *Text-Enhanced Alignment* (TEA) to address the modality imbalance problem by enhancing the correlated visual-textual knowledge. Instead of struggling with the difficulty of establishing correlation matching between infinite granularity visual semantics and partially relevant text description in existing dataset sample pairs, we take an off-the-shelf MLLM as a video narrator to generate plausible textual descriptions of the video (Fig. 1(b)). This choice of MLLM is driven by its capacity to enhance both the semantic completeness and diversity of the narratives with prompt instructions. To effectively maintain time sensibility in the generated descriptions to facilitate temporal localization, we design to generate narratives of the video at different timestamps, forming a structured text paragraph temporally aligned with the video. The structured paragraph converts the intricate and noisy video sequence data into cohesive log semantic summaries with time information, containing a comprehensive text corpus relevant to visual content, which is primed to narrow the cross-modal heterogeneous gap and aid in temporal moment localization. Then we employ a video-narrative knowledge enhancement module to merge augmented narratives with the video feature, resulting in adaptively enriched semantic-aware video representations for query localization by the multi-modal attention mechanism. Subsequently, considering the complementary semantic description provided in the structured paragraph, we also introduce a paragraph-query parallel interaction module to facilitate uni-modal video-query alignment. The semantic-enriched narratives play a crucial role in reducing the cross-modality heterogeneous gap, thereby improving the generalization and robustness of facing diverse video-text distributions.

We make three **contributions** in this work: (1) We formulate a novel paradigm called Text-Enhanced Alignment (TEA) to mitigate the modality imbalance problem and enhance generalizable moment-text associations learning in VMR. (2) We leverage an MLLM as a video narrator to construct structured textual narratives for videos. These narratives are temporally aligned with videos, serving as enriched semantic bridges to aid in video-query alignment. (3) TEA provides state-of-the-art performances on various evaluations from two popular VMR benchmarks, demonstrating the efficacy of the proposed method for enhancing generalizable visual-textual learning.

2. Related works

Video Moment Retrieval (VMR). VMR [7,19], also known as natural language video localization or video grounding, requires fine-grained temporal sensibility to associate moments and queries. Proposal-based methods [7,19,20] generate candidate video segments, aggregate all the frames with a video segment, and align them holistically with the query. Another paradigm is proposal-free boundary identification, aiming to directly regress the temporal coordinates of the target moments [21] or predict the per-frame probabilities of being the start and end points [22,23]. And some works tried to retrieve moments by the proposal-based and proposal-free strategies jointly [24–26]. All of them focused on learning fine-grained visual-textual correlation from the datasets while suffering from the modality imbalance problem.

Modality Imbalance in VMR. In existing VMR datasets, queries describing the user's moments of interest often only capture a fraction of the video content spatially and temporally. This inherent modality imbalance results in suboptimal learning of the association between moments and text. To enhance the query and promote correlated cross-modal learning, several methods [10–12] customized the rules

to very specific wording and syntax characteristics, *e.g.*, DeCo [11] decomposed and re-combined query elements in multiple granularities, potentially reducing effectiveness when confronted with unseen distributions diverging from the tailored trained data. [27] constructed a support set, considering the simultaneous presence of certain visual entities but still ignoring unrelated semantics in the vision. Some works [13,14,28] complemented the query semantics from contexted sentences, *e.g.*, TSMR [28] proposed a Teacher-Student framework that leverages multiple complementary queries to enhance textual semantics for improved language-driven moment retrieval in videos. MESM [13] discussed the word and segment-level imbalances and added prior knowledge from neighbor queries in the same video. However, they are not always realistic in real-world scenarios where the video-text annotation is limited. In contrast to the existing approaches, which rely on limited visual-textual information within datasets through various permutations or combinations, we utilize an MLLM to generate video narratives, thereby enhancing both their semantic richness and diversity, to aid in cross-modal alignment between video and query.

Large Language Models in Video Understanding. Recent developments in large language models [15] and large-scale vision-language pretraining [29,30] have underscored the abilities of MLLMs [16,31] to understand visual and textual information. Upon these, several works [32,33] trained MLLMs for video inputs but still failed to provide meaningful temporal localization predictions [18]. To pursue better fine-grained video understanding, [17] used a Q-former to fine-tune a time-sensitive MLLM with explicit time information for video reasoning. Further, [34] used GPT to generate captions for trimmed videos to facilitate text-video retrieval from a set of videos. [18] transformed videos into captions and asked GPT to predict moment boundaries directly. However, due to the information loss in video input preprocessing and limited grounding annotations, all these approaches were still struggling to get accurate moment endpoints for localization. Furthermore, some attempts [35,36] embedded vision-language pertaining knowledge in feature space for localization but still struggled with cross-modal alignment precision in heterogeneous domains. In this work, we introduce the multi-modal understanding capacity of MLLM to a VMR pipeline by generating the text narratives corresponding to videos. Different from [34], videos in VMR are untrimmed and often unscripted [26], which brings more challenges in understanding fine-grained video-text alignment knowledge and conducting temporal segment semantic searching. We carefully form a structured paragraph temporally aligned with videos at certain timestamps, bringing rich text corpora correlated with videos for cross-modal understanding.

Video Paragraph Grounding. In the context of natural language video localization, Video Paragraph Grounding (VPG) [37,38] is a related task in video-text understanding. VPG focuses on jointly localizing multiple sentences while preserving their temporal order within an untrimmed video. Given an untrimmed video and a paragraph query composed of multiple temporally ordered sentences, VPG aims to localize the temporal intervals of all events described in the paragraph simultaneously. There are two key differences between VPG and our proposed method. First, since VPG shares the same dataset sources [6,7] as VMR, it continues to suffer from modality imbalance due to the limited text corpus in existing annotations, an inherent issue that requires further investigation. Second, unlike the paragraphs in VPG, which lack temporal information, our method generates narratives that are temporally aligned with videos. These enriched narratives serve as semantic bridges, enhancing visual-textual understanding and improving video-query alignment.

3. Methods

3.1. Problem definition

Given an untrimmed video V with T duration, and a natural language query sentence Q that reflects the user's interests in specific temporal and visual parts, the object of video moment retrieval is to semantically align the query to the target video moment segment by predicting its start and end timestamps (τ^s, τ^e). It is challenging to get a semantic understanding of both visual and textual inputs and then align them to localize accurately the temporal boundaries of a certain motion behavior.

Considering the query describes only part of the video temporal information and part of the visual semantics in the target moment segment, it is hard for the model to understand visual-textual correlation knowledge comprehensively under the lack of text corpus in both semantics and syntax. In this work, we study the modality imbalance problem by proposing a *Text-Enhanced Alignment* (TEA) model (Fig. 2). TEA first generates a structured text paragraph that is temporally aligned with the input video V from an offline multi-modal large language model. The choice of MLLM is driven by its capacity to enhance both the semantic completeness and diversity of the narratives. Subsequently, TEA employs the video-narrative knowledge enhancement module to acquire more discriminative video representations enriched with augmented narrative semantics. Expanding upon this, we introduce a paragraph-query parallel interaction module to facilitate cross-modal video-query alignment, addressing the imbalance problem and cross-modal heterogeneous gap. This design aims to improve the generalization and robustness of the VMR model in managing diverse video-text distributions. Adopting the convention [23,24,35], we represent video snippet features with a pretrained CNN as $V = \{s_i\}_{i=1}^{L^v}$ composed of L^v snippets where each captures a non-overlap time period $[t^s_{s_i}, t^e_{s_i}]$, and the query sentence by the GloVe embeddings as $Q = \{w_i\}_{i=1}^{L^q}$ with L^q words.

3.2. Temporally structured text paragraph construction

Recent developments in large language models [15,16] and large-scale vision-language pretraining [29,30] have underscored the extraordinary abilities of MLLMs to understand both visual and textual information. Upon this, to mitigate the modality imbalance in video-query samples, we use one off-the-shelf MLLM as a video narrator to generate multifaceted and diverse caption descriptions related to the video content. To effectively maintain temporal sensibility in the generated descriptions for moment temporal localization, in this section, we utilize one offline MLLM to convert one video to multiple text descriptions at different timestamps and then construct a structured paragraph.

Given the input video V, we utilize a pre-trained MLLM$(\cdot)$ to transcribe the raw visual data into a list of narrative descriptions. Specifically, V is firstly sampled to image frames $\{f_i\}_{i=1}^{L^f}$ at fixed time intervals m, where L^f is the number of frames. Then we instruct MLLM$(\cdot)$ with prompt P = "This is one image frame sampled from a video. Please caption this frame in two or three sentences, to describe this frame with some details but without any analysis". to yield the text narrative of each frame f_i at a certain timepoint t_{f_i}, as follows:

$$c_i = \text{MLLM}(f_i, P), \tag{1}$$

where c_i is the response answer as the text description of f_i, capturing the video's narrative that correlates the visual semantics in a more accessible and explainable format. Then all the text narratives are concatenated in chronological order to construct a text paragraph C with time information as:

$$C = \{t_{f_i}: \ c_i\}_{i=1}^{L^f}. \tag{2}$$

In practice, the sampling rates of individual videos may vary [20,22,39] due to the diversity of video acquisition and codec processing, leading to potentially inconsistent sampling rates between videos and the constructed corresponding paragraphs. In this case, for fine-grained visual-textual semantic matching, we design to align the paragraph with the video features sequence temporally. Similar to the merge

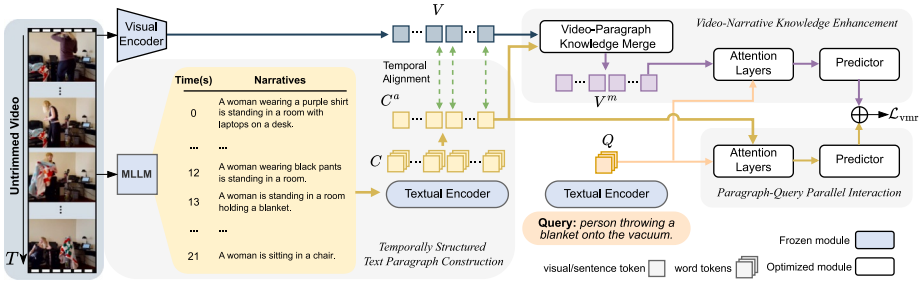

Fig. 2. An overview of our *Text-Enhanced Alignment* (TEA) model. We take an offline MLLM as a video narrator to generate a structured narrative paragraph C^a that is temporally aligned with the input video snippet feature sequences V. TEA performs video-narrative knowledge enhancement to acquire more discriminative text-enhanced video representations. Parallelly, we conduct a paragraph-query interaction module to complement context understanding and promote more generalizable predictions.

operation on video features in [22], we perform mean-pooling of the neighbor narrative features whose time points fall into the same snippet period as follows:

$$C^a = \{c_k^a\}_{k=1}^{L^v} = \text{Meanpool}(\{\text{Sent}(\mathcal{F}_T(c_m)), \text{where } t_{f_m} \in [t_{s_k}^s, t_{s_k}^e]\}), \tag{3}$$

where $\mathcal{F}_T$ is the frozen text feature extractor to get word-level embeddings and Sent$(\cdot)$ is to generate sentence-level features by averaging. After the alignment, the structured text paragraph feature C^a are semantically matched with the input video V on the time dimension. This matching process facilitates the conversion of intricate and noisy video visual sequence data into a cohesive log of semantic summaries with time sensibility, containing a valuable and comprehensive text corpus relevant to the visual content. This corpus is primed to aid in temporal moment localization. Moreover, the design of merging multiple neighbor narratives from coherent similar visual content to align with the video granularity, will also improve the robustness to resist potential noises from MLLM's output. For simplicity, we reuse C to represent the temporally aligned structured paragraph features C^a.

3.3. Video-narrative knowledge enhancement

The structured paragraph corresponding to the video contains the text narrative summary in different time periods. Due to the intrinsic characteristic of sharing the same semantic space with the query Q, Paragraph C bridges the semantic understanding between the abundant temporal visual data with abstract textual query information. In this section, we leverage the complementarity between videos and augmented text narratives to cultivate more discriminative text-enhanced video sequence representations. These representations help to narrow the cross-modal heterogeneous gap, thereby aligning video moments with the query Q.

Video-paragraph knowledge merging. To model weighted combinations of the video snippet and the text narratives, we concatenate each snippet feature s_i with the corresponding paragraph text embedding c_i on the hidden feature dimension and utilize a learnable multi-layer perceptron (MLP) to get the information-merged video snippet feature s_i^m as follows:

$$s_i^m = \text{MLP}(s_i \parallel c_i), \tag{4}$$

where $(\cdot \parallel \cdot)$ denotes the hidden-feature-wise concatenation. With the merge operation, the video feature s_i^e reduces the redundant noise in the visual input and complements the augmented text narrative information, s which is naturally more compatible with the query. Then we get the information-merged video sequence V^m:

$$V^m = \{s_i^m\}_{i=1}^{L^v}. \tag{5}$$

Query-attended knowledge enhancement. After getting the information-combined video features from the structured paragraph narratives, we facilitate knowledge enhancement by deploying attentive encoding [1,40] for both visual and textual representations to analyze the correlation among elements in both. In one attention unit $\mathcal{A}$ on sequence analysis, given a target sequence $X^t \in \mathbb{R}^{L^t \times d}$ with L^t elements and a reference $X^r \in \mathbb{R}^{L^r \times d}$ with L^r length, $\mathcal{A}(X^t, X^r)$ attends X^t using X^r as follows:

$$\begin{aligned} &\mathcal{R}(X^t, X^r) = \text{Softmax}(\text{FC}(X^t)\text{FC}(X^r)^\top / \sqrt{d}) \in \mathbb{R}^{L^t \times L^r}, \\ &\mathcal{A}(X^t, X^r) = \text{FC}(X^t + \mathcal{R}(X^t, X^r)\text{FC}(X^r)) \in \mathbb{R}^{L^t \times d}. \end{aligned} \tag{6}$$

The attention unit $\mathcal{A}(X^t, X^r)$ in Eq. (6) is parameterized by four independent fully connected (FC) layers. And we also conduct a guided mechanism [26] on video features:

$$\hat{V}^m = \text{Conv2D}(\{V^m, \{\text{Maxpool}(\{s_i^m\}_{i=1}^t)\}_{t=1}^{L^v}, \{\text{Maxpool}(\{s_i^m\}_{i=t}^{L^v})\}_{t=1}^{L^v}\}). \tag{7}$$

Then we promote self- and cross-attention for context exploration and knowledge enhancement by:

$$V^e, Q^e = \text{Attn}(\hat{V}^m, Q), \tag{8}$$

where in the Attn(X, Y) function:

$$X \leftarrow \mathcal{A}(X, X), X \leftarrow \mathcal{A}(X, Y); \quad Y \leftarrow \mathcal{A}(Y, Y), Y \leftarrow \mathcal{A}(Y, X). \tag{9}$$

After both merging and multi-modal attention operations, we combine generated narratives with video features and adaptively enrich semantic-aware video presentations for query-relevant localization.

Endpoint prediction. With the knowledge-enhanced video and query features(V^e, Q^e), our TEA model is ready to benefit existing VMR predictors. Here, we take the state-of-the-art span-based predictor [26] as an example to get per-snippet scores of being the start and end time points (p_v^s, p_v^e):

$$(p_v^s, p_v^e) = \text{Predictor}(V, Q) = \text{Softmax}(\text{LSTM}(\bar{V} \odot h)), \tag{10}$$

where

$$\begin{aligned} &h = \sigma(\text{Conv1D}(\bar{V} \parallel q)); \\ &\bar{V} = H(V, Q) = \text{FC}(V \parallel X^{v2q} \parallel V \odot X^{v2q} \parallel V \odot X^{q2v}); \text{ and} \\ &R = \text{FC}(V)\text{FC}(Q)^\top / \sqrt{d}, \ X^{v2q} = R^r Q, \ X^{q2v} = R^r R^{c\top} V; \end{aligned} \tag{11}$$

q is the sentence-level query feature by the weighted sum of words, σ is the sigmoid function, $\odot$ denotes Hadamard Product and $(\cdot \parallel \cdot)$ is concatenation. R^r, R^c are deployed row and column-wise softmax operation on R.

3.4. Paragraph-query parallel interaction

The generated structured paragraph C is a cohesive log of semantics summaries, containing a comprehensive text corpus relevant to the visual content. It also may highlight some content as a complementation for query localization. For example, if the query would like to find 'another person', the different narratives of the two people in the paragraph ('a man with pink shirt' *vs.* 'a woman with blur

dress') will provide a clear guide for the 'another' in the uni-modal understanding. Considering that the structured paragraph C shares the same semantic space with the query, and the paragraph also has temporal discrimination to guide the moment localization, in this section, we conduct a paragraph-query interaction parallelly to enhance the semantic alignment in the text-text interaction space as a complement. Specifically, given the structured paragraph C and the query Q, we first apply the similar attention interaction in Eq. (8) independently to get the attention attended paragraph and query features as follows:

$$C, Q \leftarrow \text{Attn}(C, Q). \tag{12}$$

And then the start and end point scores are calculated as:

$$(p_c^s, p_c^e) = \text{Predictor}(C, Q). \tag{13}$$

Then the final start and end point scores (p^s, p^e) are enhanced by weighting both the video-query prediction and paragraph-query prediction:

$$p^s = p_v^s + \alpha p_c^s, \ p^e = p_v^e + \alpha p_c^e, \tag{14}$$

where α is the weighted hyper-parameter.

3.5. Model training

In the training stage, we follow [26] to expand label boundaries (τ^s, τ^e) into candidate endpoint sets $(\tilde{\boldsymbol{\tau}}^s, \tilde{\boldsymbol{\tau}}^e)$ by an auxiliary proposal-level ranking and top-1 proposal selection with the boundary of (t_η^s, t_η^e). The endpoint sets are constructed as $\tilde{\boldsymbol{\tau}}^{s/e} = [\min(t_\eta^{s/e}, \tau^{s/e}), \max(t_\eta^{s/e}, \tau^{s/e})]$. More details are provided in Appendix. Then the VMR model retrieval loss is computed as:

$$\mathcal{L}_{vmr} = -\log(\sum_{i \in \tilde{\boldsymbol{\tau}}^s} p_{v,i}^s) - \log(\sum_{i \in \tilde{\boldsymbol{\tau}}^e} p_{v,i}^e), \tag{15}$$

and we highlight foreground video content by learning h in Eq. (11) as:

$$\mathcal{L}_h = \text{BCE}(y^h, h), \ y_i^h = \mathbb{1}[\min(\tilde{\boldsymbol{\tau}}^s) \le i \le \max(\tilde{\boldsymbol{\tau}}^e)]. \tag{16}$$

Then the overall loss of TEA is then formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{vmr}} + \lambda \mathcal{L}_{\text{h}}, \tag{17}$$

where λ is a hyper-parameter. TEA's overall training process is summarized in Alg. 1.

Algorithm 1 Text-Enhanced Alignment (TEA) for Video Moment Retrieval

Input: An untrimmed video V, a query sentence Q, a temporal boundary (τ^s, τ^e).
Output: An updated video moment retrieval model.
Sampling a random mini-batch of video-query pairs;
foreach video-query pair **do**
  Generate text narratives C by Eq. (1);
  Construct temporally aligned structured paragraph C^a by Eq. (3);
  Get paragraph-enhanced video representations V^e by Eq. (8);
  Conduct paragraph-query (C^a, Q) parallel interaction via Eq. (12);
  Get enhanced moment prediction scores (p^s, p^e) by (14);
end foreach
Optimize model weights by minimizing $\mathcal{L}((p^s, p^e), (\tau^s, \tau^e))$ via Eq. (17).
Updating model weights by back-propagation.

4. Experiments

Datasets. Experiments were conducted on two popular VMR benchmark datasets: **(1)** Charades-STA [7] is built upon the Charades dataset [41], which is mainly about indoor activities, for video captioning and action recognition. The work of [7] adapted the dataset to the VMR task by collecting the query annotations. **(2)** ActivityNet-Captions [6] is built on ActivityNet [42] for the dense video captioning task, which is a large-scale dataset of human activities based on YouTube videos. ActivityNet-Captions is a much larger dataset compared to Charades-STA, with more sample pairs (71.9k *vs.*16.1k) and more diverse information. Table 1 illustrates their quite different data characteristics. The average length of both moments and videos in ActivityNet-Captions are much longer than those in Charades-STA, which leads to highly varied semantics richness in vision. Regarding the query, the descriptions in Charades-STA are much shorter, many consisting only of simple subject-verb-object structures with limited vocabularies (1.3k *vs.* 12.5k), and are sometimes incomplete.

Dataset generalizability evaluation splits. To measure the visual-textual matching performance and generalization ability, there are several distinct splits according to different aspects of testing in the two benchmark datasets. CD-Test-ood [43] is proposed to effectively evaluate the video-query alignment under temporal annotation bias, especially the retrieval accuracy of specific key instances and verbs in the sentence, by introducing unseen locations in the test set. CG-Novel-word [8] is designed for testing the generalization capability of unseen words. And CG-Novel-composition [8] is to assess the compositional capability by different query wording and compositions. All three are widely used dataset splits in VMR to evaluate the cross-modal understanding quality. Given the diverse facets targeted by these splits, each poses distinct challenges.

Performance metrics. Following the convents [1,22,39] to fairly measure results, we use "IoU@m" to calculate the percentage of the top predicted moment having Intersection over Union (IoU) with ground truth larger than m, and also adopt "mIoU" to represent the average IoU over all testing samples. We report the results as $m \in \{0.5, 0.7\}$ for fair comparison following [11,39].

Implementation details. For video modality, we used the features provided by [26,39]. Specifically, for fair comparison, we used the I3D [44] features for the Charades-STA dataset, and I3D [44] and C3D [45] features for ActivityNet-Captions. GloVe embeddings [46] were utilized as the word-level text feature embeddings. Videos were downsampled to 128 frames at most by max-pooling and zero-padded the shorter ones. The dimension of all the hidden layers was fixed at 128, and the number of attention heads in Eq. (6) was 8 followed by layer normalization and 0.2 dropout rate. The frame interval m was 1 s. We took a pre-trained `LLaVA-v1.5-13b` [16] as the multi-modal large language model to generate the text descriptions. The TEA model was trained for 100 epochs by the Adam optimizer using a linearly decaying learning rate of 0.0005 and gradient clipping of 1.0 with a batch size of 16. The hyper-parameter in Eq. (17) was empirically set as $\lambda = 5$. And the weighted hyper-parameter α in (14) was set to 0.5. All experiments were implemented by PyTorch, and run on a single NVIDIA A100 40G GPU.

4.1. Comparisons with the state-of-the-arts

To validate the generality and effectiveness of our proposed TEA, we compared TEA with existing methods on all three data splits in both Charades-STA and ActivityNet-Captions datasets. The quantitative results are shown in Tables 2 and 3, respectively. One can see that TEA outperforms the SOTA methods by a significant margin on most metrics in all three tests of both Charades-STA and ActivityNet-Captions datasets. TEA can generate comparable results compared with MESE [13], whilst MESE also utilizes additional ground-truth sentences in one video to complement context. The performances in different visual-textual understanding challenges demonstrate the effectiveness of our text-enhanced alignment model. With temporally structured paragraphs, TEA can bridge the gap between video and text modalities and address the modality imbalance problem to get a more accurate

Table 1
Statistics of benchmark datasets.

Dataset	#video	#moment	Avg. len. (sec)		Avg. len. (wrd)	Vocab. Size	Query example
			Moment	Video	Query		
Charades	6672	16,128	8.1	30.6	7.2	1.3k	Person they put on their shoes.
Anet	14,926	71,957	36.2	117.6	14.8	12.5k	As the woman moving the dolphins started to swim with her, the dolphins' fins are visible as they swim up and down.

Table 2
Comparisons with SOTAs on Charades-STA. The best results are in **bold**. The gray row indicates using additional ground-truth descriptions in the same video.

Method	Year	Feature	CD-Test-ood			Feature	CG-Novel-word			CG-Novel-composition		
			IoU@0.5	IoU@0.7	mIoU		IoU@0.5	IoU@0.7	mIoU	IoU@0.5	IoU@0.7	mIoU
2D-TAN [20]	2020	I3D	35.88	13.91	34.22	I3D	29.36	13.21	28.47	30.91	12.23	29.75
LGI [47]	2020	I3D	42.90	19.29	39.43	I3D	26.48	12.47	27.62	29.42	12.73	30.09
VSLNet [22]	2020	I3D	34.10	17.87	36.34	I3D	25.60	10.07	30.21	24.25	11.54	31.43
DRN [48]	2020	I3D	31.11	15.17	23.05	–	–	–	–	–	–	–
DCM [49]	2021	I3D	45.47	22.70	40.99	–	–	–	–	–	–	–
VISA [8]	2022	–	–	–	–	I3D	42.35	20.88	40.18	45.41	22.71	42.03
Shuffling [39]	2022	I3D	46.67	27.08	44.30	–	–	–	–	–	–	–
EMB[a] [26]	2022	I3D	51.97	31.08	48.51	I3D	48.92	28.92	45.18	43.61	24.58	41.07
Primitives [10]	2023	–	–	–	–	I3D	50.36	28.78	43.15	46.54	25.10	40.00
DeCo [11]	2023	–	–	–	–	I3D	–	–	–	**47.39**	21.06	40.70
BM-DETR [50]	2023	CLIP	49.32	27.12	45.18	–	–	–	–	–	–	–
VDI [35]	2023	–	–	–	–	CLIP	46.47	28.63	41.60	–	–	–
MESM [13]	2024	–	–	–	–	I3D	50.50	33.67	46.20	46.19	26.00	41.40
TEA (Ours)	2024	I3D	**54.28**	**33.04**	**50.28**	I3D	**50.94**	**32.66**	**47.34**	45.00	**27.75**	**42.09**

[a] Denotes the reproduced results under the strictly identical setups using the code from the authors.

Table 3
Comparisons with SOTAs on ActivityNet-Captions. The best results are in **bold**.

Method	Year	Feature	CD-Test-ood			Feature	CG-Novel-word			CG-Novel-composition		
			IoU@0.5	IoU@0.7	mIoU		IoU@0.5	IoU@0.7	mIoU	IoU@0.5	IoU@0.7	mIoU
2D-TAN [20]	2020	I3D	22.01	10.34	28.31	C3D	23.86	10.37	28.88	22.80	9.95	28.49
LGI [47]	2020	I3D	23.85	10.96	28.46	C3D	23.10	9.03	26.95	23.21	9.02	27.86
VSLNet [22]	2020	I3D	20.03	10.29	28.18	C3D	21.68	9.94	29.58	20.21	9.18	29.07
DCM [49]	2021	I3D	22.32	11.22	28.08	–	–	–	–	–	–	–
VISA [8]	2022	–	–	–	–	C3D	30.14	15.90	35.13	31.51	16.73	35.85
Shuffling [39]	2022	I3D	24.57	13.21	30.45	–	–	–	–	–	–	–
EMB[a] [26]	2022	I3D	27.72	14.03	31.25	I3D	31.97	15.82	34.87	31.49	16.00	35.31
Primitives [10]	2023	–	–	–	–	C3D	30.15	14.97	32.14	30.80	15.39	33.18
DeCo [11]	2023	–	–	–	–	C3D	–	–	–	28.69	12.98	32.67
VDI [35]	2023	–	–	–	–	CLIP	32.35	16.02	34.32	–	–	–
TEA (Ours)	2024	C3D	25.77	12.56	30.72	C3D	30.59	16.48	35.20	30.98	15.37	33.90
TEA (Ours)	2024	I3D	**27.98**	**14.45**	**31.85**	I3D	**32.89**	**16.79**	**35.17**	**32.97**	**17.49**	**36.24**

[a] Denotes the reproduced results under the strictly identical setups using the code from the authors.

and generalizable cross-modal understanding. Given the larger amount of samples, richer syntax/wording expressions, and longer video/moment durations, ActivityNet-Captions poses additional challenges in video-query semantic understanding. TEA still achieves promising performances in all evaluations in Table 3. Although some methods get higher IoU@0.5 results on a certain split in one dataset by tailor-made design for specific tests, TEA can consistently achieve outstanding performances across heterogeneous datasets and tests, and promote more accurate predictions on more challenging metric IoU@0.7 and mIoU, further indicating the effectiveness of the proposed method for enhancing generalizable visual-textual learning.

4.2. Comparisons with different LLM utilization choices

As mentioned in the related works section, regarding utilizing LLM for video understanding, especially for video moment retrieval, there are broadly three strategies. One ('Captioner + LLM') is to use a captioner to convert video frames/segments into textual descriptions and then leverage the powerful ChatGPT API to process all text inputs and get the temporal boundary predictions. The second ('Video-adaptor + LLM') utilizes an adaptor (such as Q-former [17]) to transfer video input into tokens for one LLM and then ask the LLM to predict the boundaries. Furthermore, some attempts are injecting large-scale vision-language pretraining knowledge into a VMR model pipeline ('model + VLM'). The comparison results on the Charades-STA original split, which is proposed by [7] and contains 12,408/3720 video-query pairs for training and testing, are shown in Table 4. TEA used I3D [44] features in the experiments. One can see, that even though benefited from powerful ChatGPT, 'Captioner + LLM' still cannot generate very reasonable results. And it is worth noting that open-sourced LLMs failed to provide any meaningful predictions [18]. Different from all three strategies, TEA leverages one easy-acquired open-sourced LLM to generate temporally structured narrative paragraphs aligned with input videos to promote the learning of generalizable moment-text associations by video-narrative knowledge enhancement and paragraph-query interaction. The performances advanced all others and demonstrate the potential and effectiveness of TEA as a paradigm ('model + LLM')

Table 4
Comparisons with different LLM utilization choices.

Method	Type	IoU@0.5	IoU@0.7
InstructBLIP [51] + ChatGPT [52]	Captioner + LLM	7.0	1.0
VideoChat-Text (w/ChatGPT) [32]		9.0	3.0
Video-LLaMA [33]	Video-adaptor + LLM	2.7	1.2
VideoChat-Embed [32]		3.2	1.4
TimeChat [17]		46.7	23.7
VDI [35]	Model + VLM	52.32	31.37
Moment-DETR [36]		53.63	31.37
TEA (Ours)	Model + LLM	**59.52**	**41.24**

Table 5
Video-paragraph merging choices.

Implementation	IoU@m		mIoU
	0.5	0.7	
Video-only	51.97	31.08	48.51
Add	52.92	32.09	49.50
Attention	52.98	31.82	48.83
Concat+MLP	**54.28**	**33.04**	**50.28**

Table 6
Different text generator choices.

Text generator	IoU@m		mIoU
	0.5	0.7	
BLIP2 [30]	53.93	32.65	49.53
LLaVA [16]	**54.28**	**33.04**	**50.28**

Table 7
Performance of our modules on different word types.

Type	Baseline	Baseline + VN	Baseline + PQ	TEA
All (Novel word)	45.18	45.82	46.62	47.34
Novel Verb	34.65	36.44	38.24	38.45
Novel Noun	46.91	50.26	49.77	51.38

for effectively leveraging the power of LLM to aid the video moment retrieval task.

4.3. Ablation studies

In this section, we performed in-depth ablations to evaluate the efficacy of each component in TEA on the Charades-STA dataset [7] with I3D [44] features.

Component analysis. We examined the effectiveness of each proposed module in Fig. 3. The video-narrative knowledge enhancement module is denoted as 'VN' and the paragraph-query parallel interaction module is 'PQ'. We take [26] as the baseline. The video-narrative knowledge enhancement and paragraph-query parallel interaction modules proposed in TEA have their own benefits to promote more accurate visual-textual understanding. Moreover, when they are both adopted together by text-enhancing the video representations and final predictions jointly, the performance benefited more.

Different implementations of video-paragraph merging. In Eq. (4), we merge the video and corresponding paragraph knowledge with concatenation and one MLP. Here we discuss other alternative implementation choices: (a) directly sum the two features (Add): $s_i^m = \text{Add}(s_i, c_i)$; $V^m = \{s_i^m\}_{i=1}^{L^v}$; (b) apply the cross-attention like in Eq. (9) for interactions (Attention): $V^m \leftarrow \mathcal{A}(V, P)$. Table 5 presents the comparative results on CD-Test-odd, illustrating how employing a weighted combination through concatenation and MLP enables the selective merging of knowledge from both video and paragraph sources during training. This approach maintains distinguishability at each period, thereby facilitating semantic temporal moment localization.

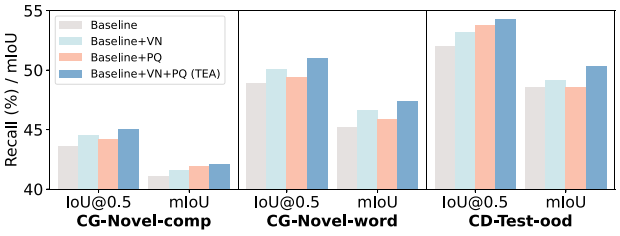

Fig. 3. Components analysis.

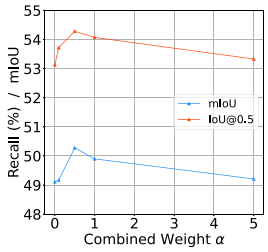

Fig. 4. Combined weight.

Different choices to generate text descriptions. In TEA, we use LLaVA [16] as the MLLM to generate text descriptions with parallel prompt inputs. There are also some valuable works about captioners without prompts. Here we select one of the representative models BLIP2 [30] (`BLIP2-pretrain-opt6.7b`) as the text generator. Table 6 shows the comparison results on CD-Test-odd, demonstrating the effectiveness and robustness of TEA to utilize various description generators to address the modality imbalance and promote more accurate retrieval. Given no prompt input in BLIP2, there is no option to adjust the granularity of the output to always maintain distinguishability at each timestamp, and the performance is not as good as promptable text generators.

Combination weight in prediction enhancement. In Eq. (14), we combine the prediction scores with the paragraph-query parallel interaction predictions with weight hyper-parameter α to add complemental context information from the text space. Here, we illustrate the hyper-parameter searching process in Fig. 4. One can see that with the increment of α, the accuracy of the predictions attains its maximum value at 0.5, and we choose $\alpha = 0.5$ as our implementation.

Further analysis of the component modules. Based on the experiments and analyses in Figs. 3 and 4, the video-narrative knowledge enhancement (VN) and paragraph-query parallel interaction (PQ) modules each contribute distinct benefits, enhancing visual-textual matching and comprehension. In this part, we further examine their roles within the overall framework. Specifically, we conduct experiments on the CG-Novel-word of Charades-STA, which introduces unseen words during evaluation. To assess the impact of our modules, we select two subsets from the test set: one comprising new verbs and the other containing novel nouns. Table 7 presents the results (mIoU) for each subset. The video-narrative knowledge enhancement module bridges generated narratives with visual information, adaptively enriching semantic-aware video representations for localization. This mechanism proves particularly beneficial for identifying entities and instances, leading to improved performance on novel nouns. Conversely, the uni-modal narrative-query matching mechanism in the PQ module emphasizes contextual coherence within descriptions, enhancing the model's understanding of temporally related information, such as the meaning of actions/verbs. When both modules are applied together,

Table 8
Results on original splits of Charades-STA and ActivityNet-Captions.

Method	Feature	Charades-STA [7]				Feature	ActivityNet-Captions [6]			
		IoU@0.3	IoU@0.5	IoU@0.7	mIoU		IoU@0.3	IoU@0.5	IoU@0.7	mIoU
VSLNet [22]	I3D	67.26	50.46	31.53	47.33	I3D	57.75	41.10	25.58	42.26
2D-TAN [20]	VGG	–	39.70	23.31	–	C3D	59.45	44.51	26.54	–
LGI [47]	I3D	72.96	59.46	35.48	51.38	C3D	58.52	41.51	23.07	41.13
DRN [48]	I3D	–	53.09	31.75	–	C3D	–	45.45	24.36	–
BPNet [25]	I3D	65.48	50.75	31.64	46.34	C3D	58.98	42.07	24.96	42.11
CPN [54]	I3D	72.94	56.70	36.62	51.85	C3D	62.81	45.10	28.10	45.70
DeNet [55]	I3D	–	59.75	38.52	–	C3D	61.93	43.79	–	–
Shuffling [39]	I3D	–	56.5	38.8	–	I3D	–	43.1	25.8	–
QD-DETR [56]	I3D	–	50.67	31.02	–	–	–	–	–	–
EMB [26]	I3D	72.50	58.33	39.25	53.09	I3D	64.13	44.81	26.07	45.59
TEA (Ours)	VGG	63.06	47.04	25.51	43.23	C3D	63.27	42.45	23.77	44.54
TEA (Ours)	I3D	73.66	59.52	41.24	53.83	I3D	64.42	45.37	27.26	46.09

Table 9
Effectiveness to other baseline. Experimental results on Charades-STA.

Method	CG-Novel-word			CG-Novel-composition			CD-Test-ood		
	IoU@0.5	IoU@0.7	mIoU	IoU@0.5	IoU@0.7	mIoU	IoU@0.5	IoU@0.7	mIoU
QD-DETR[a] [56]	44.60	22.88	40.07	38.90	19.32	35.94	44.80	20.06	40.95
QD-DETR [56] + TEA	**49.06**	**27.48**	**43.90**	**39.89**	**19.87**	**36.89**	**46.87**	**21.90**	**41.74**

[a] Denotes the re-implemented results under the same setups using the code from the authors.

Table 10
Complexity and efficiency analysis of MLLMs.

Model	GPU memory	Time (s) per caption
`LLaVA-v1.5-13b`	27716 MB	0.706
`LLaVA-v1.5-7b`	14940 MB	0.511
`BLIP2-pretrain-opt6.7b`	15907 MB	0.322

leveraging text to enhance video representations and refining final predictions, the overall performance sees further improvements.

Results on original splits. We further conduct experiments on original standard splits of Charades-STA and ActivityNet-Captions. We provide detailed results on original splits in Table 8. For a more comprehensive comparison, we also report the results using VGG [53] features on the Charades-STA dataset. In these splits, the train and test parts follow the i.i.d. assumption. We can see that our proposed method TEA outperforms the baseline [26] on all tests and achieves competitive performance over the SOTA methods, further demonstrating the effectiveness of TEA to enhance more accurate visual-textual learning and promote more precise predictions.

Effectiveness to other VMR baseline. To further evaluate the effectiveness of TEA, we embed TEA with temporally structured paragraphs and video-narrative knowledge enhancement to another popular VMR model QD-DETR [56]. Different from [26], which is a span-based model, the framework of QD-DETR is heterogeneous, which is based on the DETR-like transformer encoder and decoder. Following the implementation of QD-DETR, we use the provided concatenated Slowfast(SF) [57] and CLIP [29] features as the video representations and apply `CLIP-ViT-B/32` as the text feature extractor for the queries and structured paragraphs. Table 9 shows the results on Charades-STA. The improvements on three different visual-textual understanding challenges further demonstrated TEA's efficacy. With temporally structured paragraphs, TEA can bridge the gap between video and text modalities and is ready to benefit existing VMR methods for enhancing generalizable visual-textual learning. Moreover, the results also demonstrate our proposed method can even benefit the vision language pertaining-based (*e.g.*, CLIP) methods.

Complexity and efficiency analysis of MLLMs. Our method employs the pre-trained `LLaVA-v1.5-13b` as the default MLLM for generating text descriptions. To more accurately and comprehensively investigate the complexity and efficiency of the utilized MLLM, we evaluate the LLaVA-1.5 [16] family across different parameter scales, including `LLaVA-v1.5-13b` and `LLaVA-v1.5-7b`, as well as the BLIP2 [30] model `BLIP2-pretrain-opt6.7b`. Table 10 presents the results. Due to its larger parameter scale, `LLaVA-v1.5-13b` requires more computational resources and time to generate the desired caption narratives.

Visualization. In Fig. 5, we show examples of prediction from Charades-STA. In the CD-Test-ood split, for the query *another person walks in holding another book*, the model needs to understand who is the *another person* to get the correct prediction. Baseline fails to understand the semantics of the query and gives an incorrect answer by retrieving one similar action about the first person. When we supplement semantics by the generated structured paragraph, which provides the complementary localization hind ('A man wearing a pink shirt' *vs.* 'A woman in a blue dress'), TEA can predict a more accurate answer. In the CG-Novel-word and CG-Novel-comp splits, the generated structured paragraph helps our model understand new concepts and promotes more precise predictions.

5. Conclusion

In this work, we formulated a novel paradigm called Text-Enhanced Alignment (TEA) to solve the modality imbalance problem in VMR and enhance generalizable moment-text associations learning. We leverage an MLLM as a video narrator to construct structured textual narratives for video content. These narratives serve as enriched semantic bridges, aiding in cross-modal video-query alignment. Our TEA model provides state-of-the-art performances on various different out-of-distribution evaluations from two popular video moment retrieval benchmark datasets, demonstrating the effectiveness of the proposed method for enhancing generalizable visual-text learning.

Limitation and future works. From the experiments in Table 10, it took less than 1 s for each frame to get one narrative description from LLaVA. To further facilitate the MLLM's capability more efficiently, the train-free paradigm is a promising research direction in the VMR task. Moreover, even though we maintained the narratives' robustness through neighbor merging in Eq (3) in the Temporally Structured Text Paragraph Construction module and also demonstrated the robustness

(a) Samples from the CD-Test-ood split

(b) Samples from the CG-Novel-word split

(c) Samples from the CG-Novel-comp split

Fig. 5. Qualitative examples on Charades-STA.

of our design across different MLLMs in Table 6, designing a more intuitive standard to measure and improve the quality of MLLMs (such as regarding hallucination [58]) is a worthwhile direction to be explored in the future.

CRediT authorship contribution statement

Weitong Cai: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis. **Jiabo Huang:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Shaogang Gong:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Hailin Jin:** Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yang Liu:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. Weitong Cai wants to thank Qilei Li, Jian Hu, Shitong Sun, and Zhenyu Wang for the valuable discussions and help.

Appendix. Auxiliary proposal-level ranking in the baseline model

In Section 3.5, we follow [26] to expand label boundaries (τ^s, τ^e) into candidate endpoint sets $(\tilde{\boldsymbol{\tau}}^s, \tilde{\boldsymbol{\tau}}^e)$ by an auxiliary proposal-level ranking and top-1 proposal selection with the boundary of (t_η^s, t_η^e). Here we provide the details of this auxiliary module.

Given the knowledge-enhanced video and query features(V^e, Q^e) (simplified as (V, Q)), we generate the segment-wise features $K = \{s_j\}_{j=1}^J$ by constructing a 2D feature map [20] to represent $J = T \times T$ proposals, where the jth proposal with boundary of (t_j^s, t_j^e) is processed by max-pooling the composed frames $s_j = \text{Maxpool}(\{f_t | \forall t \in [t_j^s, t_j^e]\})$. And the segment-level representations are then encoded with temporal boundary information by the similar operation in Eq. (7):

$$\hat{K} = \text{Conv2D}(\{K, \{f_{t_j^s}\}_{j=1}^J, \{f_{t_j^e}\}_{j=1}^J\}) \in \mathbb{R}^{J \times d}, \tag{A.1}$$

and the assembled video representations $\hat{K}$ and corresponding Q are used for the cross-modal and single-modal attention the same as in Eq. (8). The two modalities features are then fused by H defined in Eq. (11) and predict the per-proposal probabilities p as:

$$p = \{p_j\}_{j=1}^J = \sigma(\text{Conv2D}(H(\hat{K}, Q))). \tag{A.2}$$

Then the ground truth overlaps y between every proposal and the manual boundary are calculated: $y = \{y_j\}_{j=1}^J$, where $y_j = \text{IoU}((t_j^s, t_j^e), (\tau^s, \tau^e))$. Then we use the binary cross entropy loss $\mathcal{L}_{ax}$ to promote cross-modal learning in the auxiliary proposal-based branch:

$$\mathcal{L}_{ax} = \text{BCE}(p, y). \tag{A.3}$$

Following [26], we expand label boundaries into candidate endpoint sets for VMR model learning by selecting the greatest (top-1) predicted score p_η in p with the boundary of (t_η^s, t_η^e), and construct the endpoint sets $\tilde{\boldsymbol{\tau}}^{s/e} = [\min(t_\eta^{s/e}, \tau^{s/e}), \max(t_\eta^{s/e}, \tau^{s/e})]$. Then the VMR model retrieval loss is calculated as:

$$\mathcal{L}_{vmr} = -\log(\sum_{i \in \tilde{\boldsymbol{\tau}}^s} p_{v,i}^s) - \log(\sum_{i \in \tilde{\boldsymbol{\tau}}^e} p_{v,i}^e), \tag{A.4}$$

where $p_{v,i}^s$ and $p_{v,i}^e$ are the per-frame (ith frame) possibilities in Eq. (10).

Then the overall loss of TEA in Eq. (17) is then updated as:

$$\mathcal{L} = \mathcal{L}_{\text{vmr}} + \mathcal{L}_{ax} + \lambda \mathcal{L}_{\text{h}}, \tag{A.5}$$

where λ is a hyper-parameter.

Data availability

The authors do not have permission to share data.

References

[1] W. Cai, J. Huang, S. Gong, Hybrid-learning video moment retrieval across multi-domain labels, in: Proceedings of the Conference on British Machine Vision Conference, BMVC, 2022.
[2] W. Cai, J. Huang, J. Hu, S. Gong, H. Jin, Y. Liu, Semantic video moment retrieval by temporal feature perturbation and refinement, in: 2024 14th International Conference on Pattern Recognition Systems, ICPRS, IEEE, 2024, pp. 1–7.
[3] J. Huang, Y. Liu, S. Gong, H. Jin, Cross-sentence temporal and semantic relations in video activity localisation, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2021, pp. 7199–7208.
[4] S. Lee, H.-I. Kim, Y.M. Ro, Text-guided distillation learning to diversify video embeddings for text-video retrieval, Pattern Recognit. 156 (2024) 110754.
[5] M. Tian, G. Li, Y. Qi, S. Wang, Q.Z. Sheng, Q. Huang, Rethink video retrieval representation for video captioning, Pattern Recognit. 156 (2024) 110744.

[6] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, J. Carlos Niebles, Dense-captioning events in videos, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 706–715.

[7] J. Gao, C. Sun, Z. Yang, R. Nevatia, Tall: Temporal activity localization via language query, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 5267–5275.

[8] J. Li, J. Xie, L. Qian, L. Zhu, S. Tang, F. Wu, Y. Yang, Y. Zhuang, X.E. Wang, Compositional temporal grounding with structured variational cross-graph correspondence learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 3032–3041.

[9] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, Uncovering hidden challenges in query-based video moment retrieval, in: Proceedings of the Conference on British Machine Vision Conference, BMVC, 2020.

[10] C. Li, Z. Li, C. Jing, Y. Jia, Y. Wu, Exploring the effect of primitives for compositional generalization in vision-and-language, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 19092–19101.

[11] L. Yang, Q. Kong, H.-K. Yang, W. Kehl, Y. Sato, N. Kobori, Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23130–23140.

[12] Z.I.A. Hakim, N.H. Sarker, R.P. Singh, B. Paul, A. Dabouei, M. Xu, Leveraging generative language models for weakly supervised sentence component analysis in video-language joint learning, 2023, arXiv preprint arXiv:2312.06699.

[13] Z. Liu, J. Li, H. Xie, P. Li, J. Ge, S.-A. Liu, G. Jin, Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Vol. 38, 2024, pp. 3855–3863.

[14] S.K. Ramakrishnan, Z. Al-Halah, K. Grauman, Naq: Leveraging narrations as queries to supervise episodic memory, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 6694–6703.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2023, arXiv preprint arXiv:2303.08774.

[16] H. Liu, C. Li, Y. Li, Y.J. Lee, Improved baselines with visual instruction tuning, 2023, arXiv preprint arXiv:2310.03744.

[17] S. Ren, L. Yao, S. Li, X. Sun, L. Hou, Timechat: A time-sensitive multimodal large language model for long video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 14313–14323.

[18] Y. Wang, Y. Yang, M. Ren, Lifelongmemory: Leveraging llms for answering queries in egocentric videos, 2023, arXiv preprint arXiv:2312.05269.

[19] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, B. Russell, Localizing moments in video with natural language, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 5803–5812.

[20] S. Zhang, H. Peng, J. Fu, J. Luo, Learning 2d temporal adjacent networks for moment localization with natural language, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Vol. 34, 2020, pp. 12870–12877.

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

[22] H. Zhang, A. Sun, W. Jing, J.T. Zhou, Span-based localizing network for natural language video localization, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, Association for Computational Linguistics, 2020, pp. 6543–6554.

[23] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, W. Lu, Interventional video grounding with dual contrastive learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 2765–2775.

[24] H. Wang, Z.-J. Zha, L. Li, D. Liu, J. Luo, Structured multi-level interaction network for video moment localization via language query, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 7026–7035.

[25] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, J. Xiao, Boundary proposal network for two-stage natural language video localization, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Vol. 35, 2021, pp. 2986–2994.

[26] J. Huang, H. Jin, S. Gong, Y. Liu, Video activity localisation with uncertainties in temporal boundary, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2022, pp. 724–740.

[27] X. Ding, N. Wang, S. Zhang, D. Cheng, X. Li, Z. Huang, M. Tang, X. Gao, Support-set based cross-supervision for video grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 11573–11582.

[28] D. Liu, X. Qu, J. Dong, G. Nan, P. Zhou, Z. Xu, L. Chen, H. Yan, Y. Cheng, Filling the information gap between video and query for language-driven moment retrieval, in: Proceedings of the ACM International Conference on Multimedia, MM, 2023, pp. 4190–4199.

[29] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning, ICML, PMLR, 2021, pp. 8748–8763.

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

[31] J. Hu, J. Lin, S. Gong, W. Cai, Relax image-specific prompt requirement in sam: a single generic prompt for segmenting camouflaged objects, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Vol. 38, 2024, pp. 12511–12518.

[32] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, Y. Qiao, Videochat: Chat-centric video understanding, 2023, arXiv preprint arXiv:2305.06355.

[33] H. Zhang, X. Li, L. Bing, Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023, arXiv preprint arXiv:2306.02858.

[34] W. Wu, H. Luo, B. Fang, J. Wang, W. Ouyang, Cap4video: What can auxiliary captions do for text-video retrieval? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 10704–10713.

[35] D. Luo, J. Huang, S. Gong, H. Jin, Y. Liu, Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23045–23055.

[36] J. Lei, T.L. Berg, M. Bansal, Detecting moments and highlights in videos via natural language queries, Proc. Conf. Neural Inf. Process. Syst. (NeurIPS) 34 (2021) 11846–11858.

[37] P. Bao, Q. Zheng, Y. Mu, Dense events grounding in video, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Vol. 35, 2021, pp. 920–928.

[38] C. Tan, Z. Lin, J.-F. Hu, W.-S. Zheng, J. Lai, Hierarchical semantic correspondence networks for video paragraph grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 18973–18982.

[39] J. Hao, H. Sun, P. Ren, J. Wang, Q. Qi, J. Liao, Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2022, pp. 130–147.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, arXiv preprint arXiv:1706.03762.

[41] G.A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, A. Gupta, Hollywood in homes: Crowdsourcing data collection for activity understanding, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2016, pp. 510–526.

[42] F. Caba Heilbron, V. Escorcia, B. Ghanem, J. Carlos Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 961–970.

[43] Y. Yuan, X. Lan, X. Wang, L. Chen, Z. Wang, W. Zhu, A closer look at temporal sentence grounding in videos: Dataset and metric, in: Proceedings of the 2nd International Workshop on Human-Centric Multimedia Analysis, 2021, pp. 13–21.

[44] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6299–6308.

[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015, pp. 4489–4497.

[46] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[47] J. Mun, M. Cho, B. Han, Local–global video-text interactions for temporal grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10810–10819.

[48] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, C. Gan, Dense regression network for video grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10287–10296.

[49] X. Yang, F. Feng, W. Ji, M. Wang, T.-S. Chua, Deconfounded video moment retrieval with causal intervention, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1–10.

[50] M. Jung, Y. Jang, S. Choi, J. Kim, J.-H. Kim, B.-T. Zhang, Overcoming weak visual-textual alignment for video moment retrieval, 2023, arXiv preprint arXiv:2306.02728.

[51] W. Dai, J. Li, D. Li, A.M.H. Tiong, J. Zhao, W. Wang, B. Li, P.N. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, in: Proceedings of the Conference on Neural Information Processing Systems, NeurIPS, 2024.

[52] OpenAI, Introducing chatgpt, 2022, URL https://openai.com/index/chatgpt/. (Accessed 11 November 2024).

[53] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[54] Y. Zhao, Z. Zhao, Z. Zhang, Z. Lin, Cascaded prediction network via segment tree for temporal video grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 4197–4206.
[55] H. Zhou, C. Zhang, Y. Luo, Y. Chen, C. Hu, Embracing uncertainty: Decoupling and de-bias for robust temporal grounding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 8445–8454.
[56] W. Moon, S. Hyun, S. Park, D. Park, J.-P. Heo, Query-dependent video representation for moment retrieval and highlight detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23023–23033.
[57] H. Fan, Y. Li, B. Xiong, W.-Y. Lo, C. Feichtenhofer, Pyslowfast, 2020, https://github.com/facebookresearch/slowfast.
[58] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, M.Z. Shou, Hallucination of multimodal large language models: A survey, 2024, arXiv preprint arXiv:2404.18930.

Weitong Cai received his B.S. and M.S. degrees from the University of Electronic Science and Technology of China in 2017 and 2020, respectively. He is currently a Ph.D. candidate in Computer Science at Queen Mary University of London, under the supervision of Prof. Shaogang Gong. He was a visiting student at National Chiao Tung University, Hsinchu. His research interests include computer vision and deep learning, currently focusing on multi-modal representation learning and video understanding.

Jiabo Huang received his Ph.D. degree in Computer Science from Queen Mary University of London under the supervision of Prof. Shaogang (Sean) Gong. He also worked closely with Dr. Xiatian Zhu from University of Surrey. His Ph.D. thesis is on unsupervised deep learning of visual representation. His current research is in Computer Vision and Machine Learning, with a focus on deep learning with insufficient human supervision, e.g. unsupervised learning, transfer learning, and vision-language modeling.

Shaogang Gong is professor of Visual Computation at Queen Mary University of London; elected a Fellow of the Royal Academy of Engineering (FREng), a Fellow of ELLIS, a Fellow of AAIA, a Fellow of the Institution of Electrical Engineers, a Fellow of the British Computer Society, a member of the UK Computing Research Committee, a Turing Fellow of the Alan Turing Institute (2018-2024), and served on the Steering Panel of the UK Government Chief Scientific Advisor's Science Review. He received the D.Phil. degree in computer vision from the Keble College, Oxford University, in 1989. He has published more than 400 research articles and seven books on topics, including Person Re-Identification, Visual Analysis of Behavior, Video Analytics for Business Intelligence, Dynamic Vision: From Images to Face Recognition, and Analysis and Modeling of Faces and Gestures. He is the inventor of 42 international patents. His research interests include computer vision, machine learning, and video analysis. He won the Institution of Engineering and Technology 2020 Achievement Medal for Vision Engineering for outstanding achievement and superior performance in contributing to public safety.

Hailin Jin received the bachelor's degree in automation from Tsinghua University, Beijing, China, in 1998, and the M.S. and D.Sc. degrees in electrical engineering from Washington University in Saint Louis, St. Louis, MO, USA, in 2000 and 2003, respectively. From Fall 2003 to Fall 2004, he was a Postdoctoral Researcher with the Computer Science Department, University of California at Los Angeles. Since 2004, he has been with Adobe Research, where he is currently a Senior Principal Scientist. He was the recipient of the Best Student Paper Award (with J. Andrews and C. Sequin) at the 2012 International CAD Conference for work on interactive inverse 3-D modeling. He is a member of the IEEE Computer Society.

Yang Liu is now a Tenure-track Assistant Professor (Ph.D. Supervisor) at Wangxuan Institute of Computer Technology, Peking University. Before joining Peking University, she was a Postdoctoral research fellow in the Visual Geometry Group (VGG) at the University of Oxford, supervised by Prof. Andrew Zisserman. She received Ph.D. and MPhil in Advanced Computer Science from the University of Cambridge, and B.Eng. in Telecommunication Engineering from Beijing University of Posts and Telecommunications (BUPT). Her research interests include computer vision, natural language processing, and machine learning, with an emphasis on how these areas can collaborate best to perform real-world tasks.