Zero-Shot Video Moment Retrieval from Frozen Vision-Language Models

Dezhao Luo[1], Jiabo Huang[1], Shaogang Gong[1], Hailin Jin[2], and Yang Liu[3][*]
[1]Queen Mary University of London
{dezhao.luo, jiabo.huang, s.gong}@qmul.ac.uk
[2]Adobe Research, [3]WICT, Peking University
hljin@adobe.com, yangliu@pku.edu.cn

Abstract

Accurate video moment retrieval (VMR) requires universal visual-textual correlations that can handle unknown vocabulary and unseen scenes. However, the learned correlations are likely either biased when derived from a limited amount of moment-text data which is hard to scale up because of the prohibitive annotation cost (fully-supervised), or unreliable when only the video-text pairwise relationships are available without fine-grained temporal annotations (weakly-supervised). Recently, the vision-language models (VLM) demonstrate a new transfer learning paradigm to benefit different vision tasks through the universal visual-textual correlations derived from large-scale vision-language pairwise web data, which has also shown benefits to VMR by fine-tuning in the target domains.

In this work, we propose a zero-shot method for adapting generalisable visual-textual priors from arbitrary VLM to facilitate moment-text alignment, without the need for accessing the VMR data. To this end, we devise a conditional feature refinement module to generate boundary-aware visual features conditioned on text queries to enable better moment boundary understanding. Additionally, we design a bottom-up proposal generation strategy that mitigates the impact of domain discrepancies and breaks down complex-query retrieval tasks into individual action retrievals, thereby maximizing the benefits of VLM. Extensive experiments conducted on three VMR benchmark datasets demonstrate the notable performance advantages of our zero-shot algorithm, especially in the novel-word and novel-location out-of-distribution setups.

1. Introduction

Given a natural video and sentence description, video moment retrieval (VMR) aims to localise a video moment based on the semantics of the sentence. This task is challenging as it requires fine-grained moment-text pairs as the learning targets [36, 41], which need to annotate not only a sentence but also the temporal position in the video corresponding to the sentence. Since assigning sentences in videos requires high accuracy, it is time-consuming and difficult to scale to web-sized datasets. To address the problem of lacking fine-grained moment-text pairs for VMR tasks, previous methods [12, 43] propose a weakly-supervised setting, which aims to learn the moment-text pairs with weak supervision such as the video-text pairs. To further reduce the reliance on annotations, recent methods propose an unsupervised VMR to localise the moment with the query chosen from a database [21] or self-generated queries [25, 32].

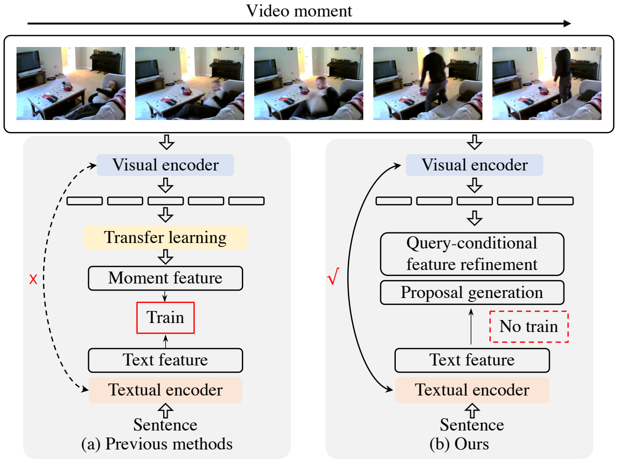

Figure 1. Unlike previous methods (a) that require additional training to fine-tune pre-trained vision-language models, our method (b) leverages pre-trained visual-textual alignment to directly predict moment-text alignment, preserving the generality of pre-trained models.

Recently, vision-language models (VLM) [6, 15, 26] have shown strong generality in VMR tasks. Specifically, PZVMR[32] and VDI [22] have explored CLIP [26] by fine-tuning the pre-learned image-text correlation for moment-

[*] Corresponding authors

text alignment learning, as shown in Fig. 1 (a). The fine-tuned moment-text alignment from limited video datasets (71K in ActivityNet-Captions [17]) is unlikely to be as generalisable as VLM pre-trained on large-scale data (e.g. 400M for CLIP [26]). Different from existing methods, we argue it can bring better generality by directly utilising pre-trained models without any additional training on domain-specific datasets. Also, image-level models are less reliable to provide a temporal understanding of the video. Even though there is still a difficulty in scaling up the fine-grained moment-text annotations, we emphasize the importance of incorporating large-scale video-text models [19, 34].

In this work, we propose a simple yet strong zero-shot VMR that fully satisfies the zero-shot requirement without the need for VMR data access. Our approach relies on the utilisation of the large-scale pre-trained video-text VLM for predicting moment-text alignments, as depicted in Fig.1 (b). The main challenge lies in the discrepancies between video-text and moment-text domains, as moment-level features necessitate the ability to discriminate between different moments within a video. This means capturing specific temporal information of the moments and understanding their various alignments to a sentence query. However, the video-text model, originally designed to retrieve textual information for the entire video, struggles to provide accurate temporal boundaries for the target moment.

To address the challenge, we adopt snippets as the fundamental units in videos and adapt the video-text model to predict the correlation between snippets and text. We recognize that each snippet is more likely to capture short-term actions, so we split the raw-query into multiple simple queries, each containing an individual action that can be better interpreted by a video snippet. To identify moment boundaries for each simple-query, we propose a conditional feature refinement module to generate boundary-aware features. Unlike previous methods [14, 25, 32] which determine boundaries based on abrupt visual changes between snippets, we argue that relying solely on spatial changes is unreliable for reflecting moment changes. Instead, we propose that the definition of suitable moment boundaries should be conditioned on the query, as different queries may emphasize different visual information. To reflect moment boundaries based on the query, we refine visual features with their context in a probability indicating how likely they are from the same moment. By suppressing visual differences within the same moment and enhancing differences between different moments, we generate boundary-aware features that are highly beneficial for VMR, even in cases where precise boundary labels are unavailable.

To generate proposals for the raw-query, we propose a bottom-up proposal generation module. We first cluster the refined snippet features into k proposals for each simple-query. Next, we perform a Cartesian product operation on the proposals obtained from all simple-queries, enumerating all possible combinations. These combinations are then merged to form final proposals for the raw-query. The scores of these proposals are determined by calculating the average snippet-text correlation using a pre-trained VLM.

Our contributions are three-folded: (1) Our zero-shot method eliminates the need for accessing the VMR data by directly applying arbitrary pre-trained VLM for moment-text alignment prediction, enabling a generalisable VMR without further training. (2) To address the discrepancies between the video-text and moment-text domains, we propose a query-conditional feature refinement module to generate boundary-aware features and a bottom-up proposal generation module to locate the final moment. (3) Our method notably outperforms existing unsupervised methods which heavily rely on human-collected videos. Importantly, it also outperforms fully-supervised methods when tested on novel-location OOD splits. Furthermore, our experiments reveal that the boundary-aware features have the potential to benefit weakly-supervised VMR where the boundary label is not provided.

2. Related Work

2.1. Video Moment Retrieval

Video moment retrieval (VMR) is a challenging task as it requires fine-grained correlation awareness between the video moment and the text.

For fully-supervised VMR, existing methods [20, 36, 40, 41] first generated visual features and textual features from pre-trained models [29, 31], then they designed a model to align the two modalities. They inevitably required a large number of annotations, which were impractical and unscalable to web-scale datasets. To alleviate the problem of fine-grained labelling, weakly-supervised methods [11, 12] proposed to learn the moment-text alignment with only a given description, relaxing VMR from marking the specific time. To further reduce the reliance on human annotations, Nam et al. [25], Wang et al. [32] and Liu et al. [21] proposed an unsupervised setting where they generated pseudo queries [25, 32] for the collected videos or chose from a query database [21]. To be noted, we regard partially zero-shot methods [25, 32] as unsupervised as they still rely on VMR-specific videos, resulting in a suboptimal solution for out-of-distribution (OOD) testing. We argue it is important to reduce the reliance on the collection of VMR data with a strict zero-shot setting for better real-world applications.

2.2. VMR with Vision-Language Pre-Training

Large-scale pre-trained vision-language models (VLM) have been explored for better video understanding. To be specific, Wang et al. [33] proposed to take the class token as an input to the sentence and build video-text align-

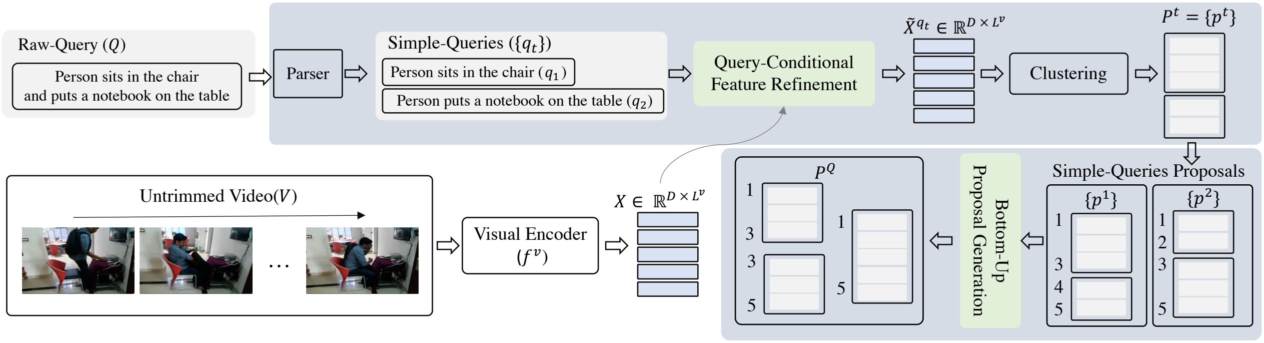
Raw-Query (Q)
Person sits in the chair and puts a notebook on the table
Parser
Simple-Queries ($\{q_t\}$)
Person sits in the chair (q_1)
Person puts a notebook on the table (q_2)
Query-Conditional Feature Refinement
$\tilde{X}^{q_t} \in \mathbb{R}^{D \times L^v}$
Clustering
$P^t = \{p^t\}$
Untrimmed Video(V)
...
Visual Encoder (f^v)
$X \in \mathbb{R}^{D \times L^v}$
P^Q
1
3
3
5
1
5
Bottom-Up Proposal Generation
Simple-Queries Proposals
$\{p^1\}$
1
3
4
5
$\{p^2\}$
1
2
3
5

Figure 2. Our framework. We first divide the raw-query Q into multiple simple-queries $\{q_t\}$, each with one verb. Then we cluster the features refined by the query-conditional feature refinement module into proposals as $\{p^t\}$ for each simple-query. In bottom-up proposal generation, we generate the final proposals P^Q for the raw-query by merging all the possible combinations from its simple-query proposals.

ment from image-text alignment for action recognition. Luo et al. [23] proposed to utilise the pre-learned VLM for video retrieval tasks. For VMR, VDI [22] proposed to inject moment information into the pre-trained text encoder and Wang et al. [32] proposed to utilise the pre-trained image-text alignment as part of the pseudo query. However, existing methods are suboptimal to leverage the generalisable visual-textual alignment learned from large-scale datasets, as they fine-tuned it for moment-text alignment on limited datasets, which is less likely to be as generalisable as the original visual-textual alignments. Moreover, video-level VLM can bring a better temporal understanding compared to image-level models, yet they are not fully explored in previous methods.

In this work, we propose a zero-shot VMR with the aim of reducing the reliance on VMR-specific dataset collection and directly exploring large-scale video-text pre-trained models for a generalisable VMR.

3. Methods

In this section, we first describe the problem and setup of our task and introduce a brief view of vision-language models (VLM), then we present the details of our approach as well as the rationale behind the design.

Problem Definition. Given an untrimmed video V that contains L^v non-overlapping snippets $V = \{s_i\}_{i=1}^{L^v}$ and a sentence query q, video moment retrieval (VMR) is performed to locate a video moment (b^v, e^v) according to the semantics of query q. The problem can be formulated as:

$$(b^v, e^v) = \mathrm{VMR}(f^v(V), f^q(q)), \tag{1}$$

where f^v denotes the *visual encoder* and its output is $X = \{\boldsymbol{x}_i\}_{i=1}^{L^v} \in \mathbb{R}^{D \times L^v}$, $\boldsymbol{x}_i$ denotes the visual representation for the i^{th} snippet s_i; f^q denotes the *textual encoder* that transfers the sentence query q into an embedding.

Setup. We detail the difference between our zero-shot setup with existing methods. Unlike fully-supervised approaches [36, 41] that rely on fine-grained moment-text pairs, weakly-supervised methods [12, 43] that depend on video-text pairs, unsupervised methods [21], or partially zero-shot methods [25, 32] that rely on human-collected videos, our approach tackles the VMR task without accessing any VMR data, such as videos, queries, or temporal annotations. Our method strictly operates in a zero-shot setting by directly leveraging pre-trained vision-language models, eliminating the requirement for a VMR dataset.

3.1. Preliminary Study of Vision-Language Models

Vision-Language Models (VLM). To learn generalisable and transferable visual and textual models, VLM train a visual encoder (ResNet [10] or ViT [5]) to map high-dimensional images/videos into a low-dimensional embedding space, and a text encoder (BERT [4]) to generate text representations from natural language. Then their correlations are learned with a contrastive loss. With a large training vision-text pair dataset (400M in CLIP [26] and 12M in InternVideo [34]), they can learn diverse visual-textual correlations that are transferable to downstream tasks.

Remarks. To study if the pre-learned visual-textual alignment is reliable for VMR, we design experiments with image-based CLIP [26] and video-based InterVideo [34] on VMR datasets: Charades-STA [7] and ActivityNet-Captions [17]. Firstly, we design a text-retrieval experiment, where the text is the query to retrieve its matched snippet. Given their foreground snippets from the groundtruth, if the VLM can allocate higher scores for the foreground other than the background, we count it as a successful retrieval. The percentage of successful sentence-retrievals out of all the samples is noted as R^t in Table 1. Also, we carry out a snippet-retrieval (R^s) experiment where the single snippet is the query to retrieve its matched sentences. In this experiment, we assess whether

Vision	Text	Charades-STA		ActivityNet-Captions	
		R^t	R^s	R^t	R^s
Random	Random	50.69	49.67	48.84	49.81
I3D	CLIP	51.19	50.77	47.39	49.20
CLIP	CLIP	70.43	63.23	69.82	65.59
InternVideo	InternVideo	**76.09**	**67.86**	**71.43**	**67.60**

Table 1. Preliminary study on the understanding of snippet-text correlation from CLIP [26] and InternVideo [34]. The number indicates the retrieval score. R^t denotes the text-retrieval and R^s the snippet-retrieval task.

the VLM can assign higher scores to the matched snippet-text pairs.

As shown in Table 1, without pre-learned alignment between I3D [3] visual encoder and CLIP [26] textual encoder, resulting in random results on both tasks. However, the pre-learned alignments from both CLIP [26] and InternVideo [34] show superior performance, indicating that they are able to understand snippet-query correlations and allocate higher scores for the matched snippet-text pairs.

3.2. Zero-Shot VMR

To address the lack of annotations and for a generalisable VMR, we propose a zero-shot method, where we take advantage of the large-scale pre-trained VLM and directly predict moment-text correlations with no additional training on VMR data such as the videos, queries or temporal annotations. As shown in Fig. 2, we break the task of locating raw-query into multiple simple-query localisations, and we design a feature refinement module to generate boundary-aware features conditioned on each simple-query.

3.2.1 Query-Conditional Feature Refinement

Based on the hypothesis that the visual feature undergoes abrupt changes at moment boundaries, previous methods have utilised CNN visual features along with hand-crafted strategies such as k-means [25] or a given threshold [14, 32] to define moment changes. However, relying solely on visual changes as indicators for moment changes is not reliable, as changes in the environment or object appearance may not necessarily correspond to moment transitions. Considering different queries may focus on different visual information to define a moment, we propose a query-conditional feature refinement module aimed at suppressing the visual differences within a moment and enhancing those between different moments. To be specific, we calculate the probability of a video snippet being in the same moment as its context snippets and refine the visual feature according to the contextual feature.

We consider the video as a series of moments, and snippets belonging to the same moment tend to exhibit similar correlation scores to the query describing the moment, while those from different moments show diverging scores. In this regard, we start by calculating their snippet-query correlation scores by f^c:

$$f^c(s, q) = \text{VLM}(s, q), \tag{2}$$

where VLM denotes the pre-trained vision-language model whose input is the snippet s and the query q.

To calculate the probability of a snippet and its context being in the same moment, we first identify the snippet that is most likely to belong to a different moment by locating the snippet with the largest snippet-query correlation difference with s:

$$\begin{aligned} D^q &= \{(f^c(s, q) - f^c(s_m, q))^2\}_{m=1}^{L^v}, \\ M^q &= \arg(\max(D^q)), \end{aligned} \tag{3}$$

where $D^q \in \mathbb{R}^{1 \times L^v}$ is the snippet-query correlation difference between s and every snippet s_m in the video, conditioned on the query q; M^q refers to the index for the snippet s_{M^q} which has the largest correlation difference with s and is considered to belong to a different moment with s. As the largest correlation difference captures the maximum disparity between moments, we use it as a metric to compute the probability of two snippets belonging to the same moment:

$$w_m^q = 1 - \frac{(f^c(s, q) - f^c(s_m, q))^2}{(f^c(s, q) - f^c(s_{M^q}, q))^2}, \tag{4}$$

where w_m^q is the probability of s_m being in the same moment with s conditioned on q.

To integrate snippets that belong to the same moment, which exhibit similar correlation scores to the sentence q, we refine the visual feature of s by:

$$\tilde{\boldsymbol{x}}^q = \boldsymbol{x} + \lambda \times X \times W^q \times \text{mask}, \tag{5}$$

where $W^q = \{w_m^q\}_{m=1}^{L^v}$ and λ is a hyper-parameter; The mask is binary values to filter context snippets. In our implementation, we only consider snippets within a distance of L^n from s and their mask values are set to 1, while snippets outside this range would have a mask value of 0.

3.2.2 Bottom-Up Proposal Generation

Since the video-text model has difficulties in generating fine-grained boundaries within the video, we adapt it for snippet-text correlation prediction. As snippets are the fundamental units of a video and are more likely to display short-term actions, we relax the raw-query retrieval task to multiple simple-query retrievals. The motivation behind this approach is to divide complex actions into individual actions to better leverage the video-text model. To be specific, we utilise a language parsing tool to parse the raw-query Q into several simple-queries q_t by extracting the verbs and their corresponding words:

$$\text{Parser(Q)}^{1} = \{q_t\}_{t=1}^{L^Q}, \tag{6}$$

where L^Q is the number of simple-queries extracted from the query Q. Then the visual feature X is refined to $\tilde{X}^{q_t} = \{\tilde{\boldsymbol{x}}_i^{q_t}\}_{i=1}^{L^v}$, conditioned on each simple-query q_t with Eq. (5). Then we cluster the features into k proposals:

$$P^t = \{p_n^t\}_{n=1}^{k}, \tag{7}$$

where P^t refers to the proposal list generated for the t^{th} simple-query q_t.

For proposal scoring, unlike previous methods [22, 32] fine-tuning vision-language models to learn moment-text alignment, we simply utilise the pre-trained video-text alignment to predict the moment-text alignment with an averaging approach:

$$\tilde{C}(p^t) = \frac{\sum_{i=b^{p^t}}^{e^{p^t}} f^c(s_i, q_t)}{e^{p^t} - b^{p^t}}, \tag{8}$$

where $\tilde{C}(p^t)$ denotes the correlation score between proposal p^t and simple-query q_t; b^{p^t}/e^{p^t} denotes the beginning and ending snippet index of p_t; The snippet-query correlation f^c is calculated as Eq. (2).

To generate the final proposals for the raw-query Q, we propose a bottom-up strategy to merge all the results obtained from simple-queries. To be specific, we first generate proposals $\{P^t\}$ for every simple-queries $\{q_t\}$ using Eq. (7). Then we use a Cartesian product to enumerate all the possible combinations of these proposals. Finally, we take the union of these proposals from the Cartesian product to generate final proposals and their corresponding scores are averaged from simple-query proposal scores:

$$\begin{aligned}
P^Q &= \cup(\{P^1 \times \cdots \times P^{L^Q}\}) \\
&= \{\{p^1 \cup \cdots \cup p^{L^Q}\} | p^1 \in P^1 \wedge \cdots \wedge p^{L^Q} \in P^{L^Q}, \\
&\quad IoU(p^1, \cdots, p^{L^Q}) > 0\}, \\
C^Q &= \{\{\frac{\sum(\tilde{C}(p^1), \cdots, \tilde{C}(p^{L^Q}))}{L^Q}\}| \\
&\quad p^1 \in P^1 \wedge \cdots \wedge p^{L^Q} \in P^{L^Q}, IoU(p^1, \cdots, p^{L^Q}) > 0\},
\end{aligned} \tag{9}$$

where "×" denotes the Cartesian product; P^Q denotes the proposal list for the raw-query Q and C^Q are their corresponding scores. Fig. 2 demonstrates an example of k =2 and L^Q = 2. The overall process is summarised in Alg. 1.

[1] Allennlp:https://allenai.org/allennlp

```
Algorithm 1 Zero-Shot VMR
Input: Untrimmed videos $V$, A query sentence $Q$, A visual $f^v$ encoder, A pre-trained model VLM.
Output: Video moment candidates $(P^Q)$ with their scores $(C^Q)$.
Compute the features of videos $X$ by $f^v$;
Generate simple-queries $\{q_t\}_{t=1}^{L^Q}$ (Eq. (6)) ;
$A \leftarrow \{\}$
for $t \leftarrow 1$ to $L^Q$ do
    Calculate the context probability conditioned on $q_t$ as $W^{q_t} = \{w_m^{q_t}\}$, with the VLM (Eq. (4));
    Generate the boundary-aware features from $X$ to $\tilde{X}^{q_t}$ (Eq. (5));
    Generate proposals $P^t = \{p^t\}$ by clustering the boundary-aware feature (Eq. (7));
    Calculate the correlation score between the proposal and query $\tilde{C}(p^t)$ (Eq. (8));
    $A$.append($P^t$)
end for
Unify the Cartesian product of $A$ to generate $P^Q$ and average their corresponding scores as $C^Q$(Eq. (9));
```

4. Experiments

With the aim of fully exploring the generalisable video-text alignment from large-scale pre-trained models, we propose to directly utilise their visual and textual encoder for video moment retrieval (VMR) without any further training. To validate the generality and effectiveness of our method, we compare with existing methods on both out-of-distribution (OOD) and independent and identically distributed (IID) data splits. In this section, we first explain the implementation details and then report our results in comparison with recent methods with a specific focus on unsupervised methods where no annotation is required. Finally, we carry out ablation studies to evaluate each module.

4.1. Experimental Settings

4.1.1 Dataset

Charades-STA [7] is built upon the Charades dataset [30] for action recognition and localisation. Gao et al. [7] adapt the Charades dataset to VMR by collecting the query annotations. The Charades-STA dataset contains 6,670 videos and involves 16,124 queries. The average duration of the videos is 30.59 seconds and the moment has an average duration of 8.09 seconds. There are 37 long-moments ($L_{mom}/L_{vid}\geq$ 0.5) out of 16,124 in this dataset.

ActivityNet-Captions [17] is collected for video captioning task from ActivityNet [2] where the videos are associated with 200 activity classes. The ActivityNet-Captions dataset consists of 19,811 videos with 71,957 queries. The average duration of the videos is around 117.75 seconds and

Method	Year	Setup	Charades-STA						ActivityNet-Captions*					
			OOD-1			OOD-2			OOD-1			OOD-2		
			0.5	0.7	mIoU	0.5	0.7	mIoU	0.5	0.7	mIoU	0.5	0.7	mIoU
LGI [24]	2020	Fully -Supervised	42.1	18.6	41.2	35.8	13.5	37.1	16.3	6.2	22.2	11.0	3.9	17.3
CMI[37]	2020		30.4	16.4	30.3	28.1	13.6	29.0	12.3	5.2	19.1	10.0	4.2	16.8
2D-TAN [41]	2020		27.1	13.1	25.7	21.1	8.8	22.5	16.4	6.6	23.2	11.5	3.9	19.4
DCM[39]	2021		44.4	19.7	42.3	38.5	15.4	39.0	18.2	7.9	24.4	12.9	4.8	20.7
MMN† [36]	2022		31.6	13.4	33.4	27.0	9.3	30.3	20.3	7.1	26.2	14.1	5.2	20.6
VDI† [22]	2023		25.9	11.9	26.7	20.8	8.7	22.0	20.9	7.1	27.6	14.3	5.2	23.7
CNM† [42]	2022	Weakly -Supervised	9.9	1.7	21.6	6.1	0.5	16.6	6.1	0.4	21.0	2.5	0.1	16.8
CPL† [43]	2022		29.9	8.5	32.2	24.9	6.3	30.5	4.7	0.4	21.1	2.1	0.2	17.7
PSVL†[25]	2021	Un -Supervised	3.0	0.7	8.2	2.2	0.4	6.8	-	-	-	-	-	-
PZVMR [32]	2022		-	8.6	25.1	-	6.5	28.5	-	4.4	**28.3**	-	2.6	19.1
Ours	2023	Zero-Shot	**40.3**	**18.2**	**38.2**	**38.9**	**17.0**	**37.8**	**18.4**	**6.8**	21.1	**18.6**	**7.4**	**20.6**

Table 2. Novel-location OOD testing. ActivityNet-Captions* denotes the datasets removed from the long-moment for fair comparisons [39]. "†" denotes our implementation with their released models. "-" denotes the model is not available, and the performance is not reported. Methods using pre-trained VLM alignments are underlined.

Method	Setup	Charades-STA		ActivityNet-Captions	
		0.5	0.7	0.5	0.7
LGI	Fully -Supervised	26.48	12.47	23.10	9.03
VISA		42.35	20.88	30.14	15.90
VDI		46.47	28.63	32.35	16.02
CNM†	Weakly- Supervised	32.52	14.82	23.11	10.21
CPL†		45.90	22.88	21.71	9.08
Ours	Zero-Shot	**45.04**	**21.44**	**24.57**	**10.54**

Table 3. Comparison with methods on the novel-word split [18]. "†" denotes the same with Table 2.

Method	Setup	Charades-STA		ActivityNet-Captions	
		0.5	0.7	0.5	0.7
DCM	Fully -Supervised	45.47	22.70	22.32	11.22
Shuffling		46.67	27.08	24.57	13.12
CNM†	Weakly- Supervised	30.61	15.23	12.89	4.06
CPL†		41.09	21.91	8.47	1.67
Ours	Zero-Shot	**40.27**	**16.27**	**19.40**	**7.85**

Table 4. Comparison with methods on the novel-distribution split [9]. "†" denotes the same with Table 2.

the moment has an average duration of 37.14 seconds. For this dataset, there are 15,736 long-moments out of 71,957. **TaCoS** [27] consists of 127 videos from MPIICooking [28]. It is comprised of 18,818 video-text pairs of cooking activities in the kitchen annotated by Regneri et al. [27].

4.1.2 Implementation Details

Evaluation Metrics. We take "R@n, IoU = μ" and "mIoU" as the evaluation metrics, which denotes the percentage of queries having at least one result whose intersection over union (IoU) with ground truth is larger than μ in top-n retrieved moments. "mIoU" is the average IoU over all testing samples. We report the results as n $\in \{1\}$ with μ $\in \{0.1, 0.3, 0.5, 0.7\}$. Following DCM [39], we collect ActivityNet-Captions* where long-moments (L_{mom}/L_{vid} $\geq$ 0.5) are removed from ActivityNet-Captions.

Hyper-Parameters. For feature extraction on video snippet, we apply I3D [3] on Charades-STA and C3D [31] on ActivityNet-Captions. For snippet-text correlation, we apply the pre-trained video-level InternVideo model [34].

For feature refinement, the context distance (L^n) is set to be 2, the λ is 0.5. We select k-means for clustering and the k to be 6. We sample the length (L^v) of each video as 32.

4.2. Comparison with the SOTAs

As a strict zero-shot VMR method requires no access to VMR data, we focus on comparison with existing unsupervised methods [21, 25, 32] in both OOD and IID testing.

4.2.1 Novel-Location OOD Testing

We first carry out experiments on a novel-location OOD scenario, where the location of the moment is different in the training set. Following DCM [39], we add a randomly generated video with p seconds in the beginning of the testing video to modify the location of moment annotations. OOD-1 and OOD-2 refers to $p \in \{10, 15\}$ for Charades-STA and $p \in \{30, 60\}$ for ActivityNet-Captions*. As shown in Table 2, we compare with existing methods across different setups. For Charades-STA, we achieve 40.3%/18.2% for OOD-1, outperforming unsupervised methods with a significant margin. For ActivityNet-

Method	Year	Setup	Charades-STA				ActivityNet-Captions			
			0.3	0.5	0.7	mIoU	0.3	0.5	0.7	mIoU
2D-TAN[41]	2020	Fully -Supervised	57.31	45.75	27.88	41.05	60.32	43.41	25.04	42.45
MMN[36]	2022		65.43	53.25	31.42	46.46	64.48	48.24	29.35	46.61
VCA [35]	2021	Weakly -Supervised	58.58	38.13	19.57	38.49	50.45	31.00	-	33.15
CNM [42]	2022		60.04	35.15	14.95	38.11	55.68	33.33	13.29	37.55
CPL [43]	2022		65.99	49.05	22.61	43.23	55.73	31.37	13.68	36.65
Huang et al. [13]	2023		69.16	52.18	23.94	45.20	58.07	36.91	-	41.02
PSVL [25]	2021	Un -Supervised	46.47	31.29	14.17	31.24	44.74	30.08	14.74	29.62
Gao and Xu [8]	2021		46.69	20.14	8.27	-	46.15	26.38	11.64	-
DSCNet [21]	2022		44.15	28.73	14.67	-	47.29	28.16	-	-
PZVMR [32]	2022		46.83	33.21	18.51	32.62	45.73	31.26	**17.84**	30.35
Kim et al. [16]	2023		52.95	37.24	19.33	36.05	47.61	**32.59**	15.42	31.85
Ours	2023	Zero-Shot	**56.77**	**42.93**	**20.13**	**37.92**	**48.28**	27.90	11.57	**32.37**

Table 5. IID testing results on Charades-STA and ActivityNet-Captions. "-" denotes the same with Table 2. Methods using pre-trained VLM alignments are underlined.

Captions, we follow DCM [39] to remove long-moments, noted as ActivityNet-Captions*, for fair comparisons. As one can see from Table 2, our method obtains 18.4%/6.8% on OOD-1, which reaches the SOTA performance among existing unsupervised methods.

It is worth noting that we demonstrate superior performance on OOD-2 over fully-supervised methods. We argue that models trained with a moment-location biased dataset (21.87% long-moments in ActivityNet-Captions) are inferior to be applied to novel-location OOD scenarios, highlighting their limitations in real-world applications.

4.2.2 Novel-Word OOD Testing

To further demonstrate our generality, we carry out testing on the novel-word split released by VISA [18], where the testing split contains novel-words not seen in the training. As shown in Table 3, with novel-word testing, we achieve the performance of 24.57%/10.54% for ActivityNet-Captions, outperforming the existing weakly-supervised models [42, 43].

4.2.3 Novel-Distribution OOD Testing

In Table 4, we further carry out testing on the novel-distribution split released by Shuffling [9], where they shuffle the moment and change the distribution of moments in the testing split. One can see from Table 4, we achieve the performance of 19.40%/7.85% for ActivityNet-Captions, outperforming existing weakly-supervised models [42, 43].

4.2.4 IID Testing

To evaluate the effectiveness of our method, we also conduct experiments on IID testing, where the training and testing split share independent and identical distribution.

Method	Setup	0.1	0.3	0.5
MCN [1]	Fully-Supervised	14.42	-	5.58
CTRL [7]		24.32	18.32	13.30
QSPN [38]		25.31	20.15	15.32
2D-TAN [41]		47.59	37.29	25.32
MMN [36]		51.39	39.24	26.17
Ours	Zero-Shot	**27.49**	**11.20**	**5.57**

Table 6. Comparison on the original split of TaCoS.

Method	VLM	Charades-STA		ActivityNet-Captions*	
		0.5	0.7	0.5	0.7
PZVMR	CLIP	-	6.5	-	2.6
Ours	CLIP	25.9	9.9	15.7	5.3
Ours	InterVideo	**38.9**	**17.0**	**18.6**	**7.4**

Table 7. Comparison between methods using pre-trained visual-textual alignments from different VLM.

As shown in Table 5, we outperform unsupervised methods [16, 25, 32] on Charades-STA, whilst we don't require any training or access to VMR dataset. For ActivityNet-Captions, we argue existing methods benefit from the moment-location bias on this dataset. Moreover, we carry out experiments on TaCoS in Table 6.

4.3. Ablation Study

We report ablations on Charades-STA and ActivityNet-Captions with novel-location OOD-2 testing.

4.3.1 Vision-Language Model Ablation

In this subsection, we compare the option of VLM between CLIP [26] and InterVideo [34]. As shown in Table 7, our proposed method demonstrates better generality from the

Method	Charades-STA			ActivityNet-Captions		
	0.3	0.5	0.7	0.3	0.5	0.7
w/o QC-FR	58.33	37.28	14.11	39.63	24.29	9.70
w/o BU-PG	60.01	**39.01**	16.68	38.95	23.83	9.86
Ours	**60.22**	38.92	**16.96**	**40.92**	**25.70**	**10.56**

Table 8. Ablation study of query-conditional feature refinement (QC-FR) and bottom-up proposal generation (BU-PG).

Method	0.3	0.5	0.7	mIoU
w/o QC-FR	65.99	49.05	22.61	43.23
w QC-FR	**66.91**	**50.85**	**24.00**	**44.00**

Table 9. Comparison of weakly-supervised CPL [43] performance with and without our QC-FR module on Charades-STA.

Method	$L^Q = 1$			$L^Q \geq 2$		
	0.3	0.5	0.7	0.3	0.5	0.7
w/o BU-PG	**32.80**	**18.59**	7.79	43.02	27.18	11.02
w/ BU-PG	32.63	18.58	**7.82**	**45.85**	**29.36**	**12.16**

Table 10. Performance on ActivityNet-Captions for raw-queries with different numbers of simple-queries L^Q.

pre-trained CLIP than previous PZVMR [32]. Furthermore, with a better understanding of the temporal information from InterVideo, our method achieves further improvement.

4.3.2 Component Ablation

In this subsection, we evaluate the effectiveness of our components. As shown in Table 8, without our proposed query-conditional feature refinement (QC-FR) module, there is a performance drop on both datasets. Also, to validate the advantages of the generated boundary-aware feature in scenarios where boundary labels are absent, we apply the query-conditioned feature refinement module (QC-FR) in weakly-supervised CPL [43] on IID testing. One can see from Table 9, by refining the feature with QC-FR, we observe performance gains of 1.80% and 1.39% on IoU = 0.5/0.7.

For bottom-up proposal generation (BU-PG), our method shows enhanced performance when evaluated with ActivityNet-Captions in Table 8. To further demonstrate the benefits of BU-PG for complex-queries, we present the performance based on the number of simple-queries for each raw-query (L^Q). It can be observed in Table 10 that BU-PG achieves superior performance when dealing with complex-queries that consist of more simple-queries ($L^Q \geq 2$).

4.3.3 Hyper-Parameter Ablation

Method	0.3	0.5	0.7	mIoU
Random	45.03	20.37	6.75	27.35
Sliding Window	49.54	38.41	14.62	33.23
Abrupt Change	20.97	9.95	2.80	14.11
K-Means (Ours)	**60.22**	**38.92**	**16.96**	**37.80**

Table 11. Ablation study of clustering method on Charades-STA.

k	0.3	0.5	0.7	mIoU
5	59.35	35.51	13.17	36.47
6	**60.22**	38.92	**16.96**	**37.80**
7	58.39	**39.01**	16.83	37.65

Table 12. Ablation study of k-means on Charades-STA.

λ	L^n	0.3	0.5	0.7	mIoU
0.1	2	58.73	38.39	15.75	37.49
0.5		**60.22**	38.92	**16.96**	**37.80**
1		59.52	**39.06**	16.48	37.19
0.5	1	59.81	38.09	16.75	37.65
	3	59.68	38.55	16.86	37.48

Table 13. Ablation study of λ and L^n on Charades-STA.

For hyper-parameters, we report the ablation on Charades-STA. We ablate the option of clustering method and select k-means as shown in Table 11. Then we take 6 as the value of "k" for k-means as shown in Table 12. The ablations of "λ" and "L^n" for feature refinement are shown in Table 13.

5. Conclusion and Future Work

In this work, we approach the video moment retrieval (VMR) task by adapting the generalisable video-text pre-trained models without requiring additional training on the target domain. To address the discrepancy between video-text and moment-text domains, we propose a query-conditional proposal generation module to generate boundary-aware features and a bottom-up proposal generation module for complex-query localisation. Superior performances on OOD testing demonstrate our method can extract generalisable moment-text alignments from pre-trained video-text alignments. For future work, one important direction is to address the challenge of understanding the temporal relationship between individual actions which is not captured by video-text pre-training models.

Acknowledgements

This work was supported by the China Scholarship Council, the Alan Turing Institute Turing Fellowship, Veritone, Adobe Research and Zhejiang Lab (NO. 2022NB0AB05).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 7

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4, 6

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1*, pages 4171–4186, 2019. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[6] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 1

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 3, 5, 7

[8] Junyu Gao and Changsheng Xu. Learning video moment retrieval without a single annotated video. *IEEE TCSVT*, 32 (3):1646–1657, 2021. 7

[9] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *ECCV*, pages 130–147. Springer, 2022. 6, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[11] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021. 2

[12] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 724–740. Springer, 2022. 1, 2, 3

[13] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *CVPR*, pages 18908–18918, 2023. 7

[14] Mihir Jain, Amir Ghodrati, and Cees GM Snoek. Actionbytes: Learning from trimmed videos to localize actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1171–1180, 2020. 2, 4

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[16] Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Language-free training for zero-shot video grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2539–2548, 2023. 7

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2, 3, 5

[18] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, pages 3032–3041, 2022. 6, 7

[19] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *CVPR*, pages 23119–23129, 2023. 2

[20] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. *arXiv preprint arXiv:2201.00454*, 2022. 2

[21] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1683–1691, 2022. 1, 2, 3, 6, 7

[22] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Towards generalisable video moment retrieval: Visual-dynamic injection to image-text pre-training. In *CVPR*, pages 23045–23055, 2023. 1, 3, 5, 6

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819, 2020. 6

[25] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1470–1479, 2021. 1, 2, 3, 4, 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 7

[27] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. doi: 10.1162/tacl_a_00207. URL https://aclanthology.org/Q13-1003. 6

[28] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012. 6

[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 2

[30] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526. Springer, 2016. 5

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2, 6

[32] Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. Prompt-based zero-shot video moment retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 413–421, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[33] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[34] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 3, 4, 6, 7

[35] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACM MM*, pages 1459–1468, 2021. 7

[36] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, volume 36, pages 2613–2623, 2022. 1, 2, 3, 6, 7

[37] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, volume 34, pages 12386–12393, 2020. 6

[38] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, volume 33, pages 9062–9069, 2019. 7

[39] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 6, 7

[40] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.585. 2

[41] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877, 2020. 1, 2, 3, 6, 7

[42] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3517–3525, 2022. 6, 7

[43] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 6, 7, 8