Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language

Xiang Fang[1]*, Daizong Liu[2]*, Wanlong Fang[3,1]*, Pan Zhou[1]†, Zichuan Xu[4], Wenzheng Xu[5], Junyang Chen[6], Renfu Li[7]

[1]Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science of Technology Wuhan, China
[2]Peking University
[3]Henan University
[4]Dalian University of Technology
[5]Sichuan University
[6]Shenzhen University
[7]Huazhong University of Science and Technology
xfang9508@gmail.com, dzliu@hust.edu.cn, wanlongfang@gmail.com, panzhou@hust.edu.cn, z.xu@dlut.edu.cn, wenzheng.xu@scu.edu.cn, junyangchen@szu.edu.cn, renfu.li@hust.edu.cn

Abstract

Given an untrimmed video and a sentence query, video moment retrieval using language (VMR) aims to locate a target query-relevant moment. Since the untrimmed video is overlong, almost all existing VMR methods first sparsely downsample each untrimmed video into multiple fixed-length video clips and then conduct multi-modal interactions with the query feature and expensive clip features for reasoning, which is infeasible for long real-world videos that span hours. Since the video is downsampled into fixed-length clips, some query-related frames may be filtered out, which will blur the specific boundary of the target moment, take the adjacent irrelevant frames as new boundaries, easily leading to cross-modal misalignment and introducing both boundary-bias and reasoning-bias. To this end, in this paper, we propose an efficient approach, SpotVMR, to trim the query-relevant clip. Besides, our proposed SpotVMR can serve as plug-and-play module, which achieves efficiency for state-of-the-art VMR methods while maintaining good retrieval performance. Especially, we first design a novel clip search model that learns to identify promising video regions to search conditioned on the language query. Then, we introduce a set of low-cost semantic indexing features to capture the context of objects and interactions that suggest where to search the query-relevant moment. Also, the distillation loss is utilized to address the optimization issues arising from end-to-end joint training of the clip selector and VMR model. Extensive experiments on three challenging datasets demonstrate its effectiveness.

Introduction

As an emerging and challenging cross-modal task, video moment retrieval using language (VMR) (Anne Hendricks et al. 2017; Gao et al. 2017) has drawn increasing attention in recent years due to its various applications, such as video understanding (Liu et al. 2023h, 2020, 2021b, 2023b, 2022a, 2021a, 2023g,a, 2022c, 2023c,d, 2022b; Fang et al. 2020, 2021a,b) and temporal action localization (Zhang et al. 2020b; Fang and Hu 2020; Fang et al. 2022, 2023a,b,c; Ji et al. 2023e, 2018, 2023g,f,d,c, 2021, 2020, 2019). As shown in Figure 1(a), the VMR task targets locating a video

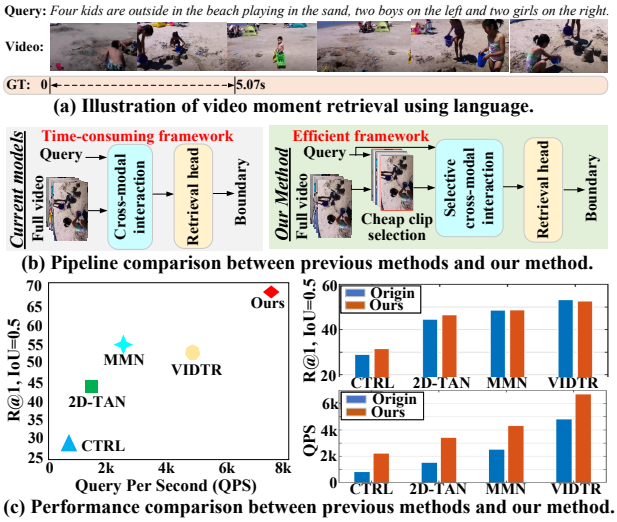

Figure 1: (a) Example of the VMR task, where GT means the ground-truth boundary. (b) Pipeline comparison between previous models and our model. Previous models trim a long video into multiple fixed-length clips and perform costly processing of every clip. These processed clips are fed to the VMR model. We propose an efficient clip selection approach that adaptively spots query-relevant clips quickly, and selectively processes these clips to serve as inputs to the VMR model. (c) Performance comparison with state-of-the-art VMR works on Charades-STA. Best viewed in color.

*These authors contributed equally.
†Corresponding Author.

moment that semantically corresponds to a given language query from a long untrimmed video. Most of the video contents are query-irrelevant, where only a short video segment matches the query. It is substantially more challenging since a well-designed method needs to not only model the complex cross-modal interaction between videos and queries, but also capture complicated context information for cross-modal semantics alignment. Not only does it require recognizing objects and activities, but also identifying which visual content is sufficient to retrieve the accurate moment expressed in free-form natural language, accounting for the fact that the accurate moment may occupy only a tiny portion of the entire video. Moreover, all this must be done in a scalable manner, given that a long untrimmed video (*e.g.*, surveillance video and live video) will ultimately span hours, days or more. In practice, VMR is an extremely challenging task because the desired model should (i) cover various moment lengths in multiple scenarios; (ii) bridge the semantic gap between different modalities (video and query); (iii) understand the semantic details of different modalities to extract modal-invariant features for optimal retrieval.

Most previous VMR works (Zheng et al. 2023; Shen et al. 2023; Yang et al. 2022; Dong et al. 2022a,b,c, 2023b,a; Sun et al. 2023; Ma et al. 2020; Liu et al. 2018, 2023f; Ge et al. 2019; Zhang et al. 2019a; Qu et al. 2023, 2021; Wen et al. 2023b, 2021, 2023a) are under fully-supervised setting, where each frame is manually labeled as query-relevant or not. Therefore, the main challenge in such a setting is how to align multi-modal features well to predict precise moment boundaries. These fully-supervised approaches can be divided into two categories: 1) Top-down approaches (Anne Hendricks et al. 2017; Chen et al. 2018; Zhang et al. 2019b, 2020b): These methods integrate sentence information with each fine-grained video clip unit, and predict the similarity scores of candidate segment proposals by gradually merging the fusion feature sequence over time. The best proposal with the highest score is selected as the predicted segment. 2) Bottom-up approaches (Chen et al. 2020; Mun, Cho, and Han 2020; Zhang et al. 2020a): These methods leverage the interaction between video and sentence to directly regress the start and end boundary frames of the target segment or predict boundary probabilities frame-wisely. The predicted segment is obtained through post-processing steps that group or aggregate all frame-wise predictions. Obviously, the frame-based annotation is very time-consuming, which will limit the applications of these methods. Although the above two types of works have achieved significant performances, they still suffer from the redundant proposal generation/matching process (top-down) and complex post-processing steps (bottom-up) to refine the grounding results.

Although the above VMR methods have made exciting headway, they neglect the practical scaling issue: they extract expensive spatio-temporal visual features for densely sampled clips throughout the long video, which spends the most computational cost of a VMR model. Such a time-consuming approach becomes intractable as the video duration grows, especially for real-time applications like surveillance video and live video, where the constrained on-board computation severely limits the applications of previous heavy-weight models. Fortunately, we can notice that (i) not all parts of the video are useful for reasoning about a given query, and (ii) there are high-level visual semantics about objects and activities that could steer our attention toward where to retrieve. For example, given a query ("A person is hitting the sides of their bed with the palm of their hand."), we can ignore video clips recorded in some irrelevant scenes other than the bedroom. Notably, these associations cannot be neatly enumerated, however, given the free-form nature of the queries. In the query ("Four kids are outside in the beach playing in the sand, two boys on the left and two girls on the right"), the model reasoning will become more complex since we need to reason about all persons in the beach (two boys and two girls), and identify their genders and position. Thus, we should learn query-conditioned priors that can use such high-level semantics to narrow down our task.

In this paper, considering that the target moment is often located near key frames or key clips, we build upon these intuitions to propose a novel and effective approach to make a given VMR method more efficient and effective. The idea is to preview the video using cheap indexing features, intelligently select a small subset of *query-relevant* clips, and only use these clips for the target moment retrieval. This can cut down computational costs without sacrificing model performance. To tackle this challenging setting with long-form videos and language queries, we design a novel clip selection architecture, which introduces a cross-modal transformer to recursively preview the video and identify some query-relevant clips. In the VMR task, the key visual features include three kinds: background feature to locate the place, appearance feature to detect the target instance, and the motion feature to recognize the activity mentioned in the query. Thus, we further design the above three kinds of semantic-indexing features that capture video context about **b**ackground features, **a**ppearance features, and **m**otion features (*BAM*). With the high-level BAM visual features as the index, we can recursively update the selected clips with the help of query features. Further, we design teacher-based distillation losses to optimize the cross-modal interaction.

To sum up, our main contributions are as follows:

- In this paper, we target a novel and efficient clip selection approach for VMR, which first previews the video using cheap indexing features, then selects a small subset of query-relevant clips, and finally only uses these selected clips for the final moment retrieval.
- We propose an effective clip selection module by designing three high-level features (BAM features) as the semantic indexer. Then, an adaptive clip update strategy can update selected clips with a feature distillation loss as supervision during each iteration.
- Our experiments on three popular yet challenging benchmarks demonstrate that our approach is more effective and efficient than state-of-the-art methods.

Related Works

Most existing VMR methods (Zhang et al. 2019b; Qu et al. 2020; Liu and Hu 2022; Liu, Qu, and Zhou 2021; Liu, Qu,

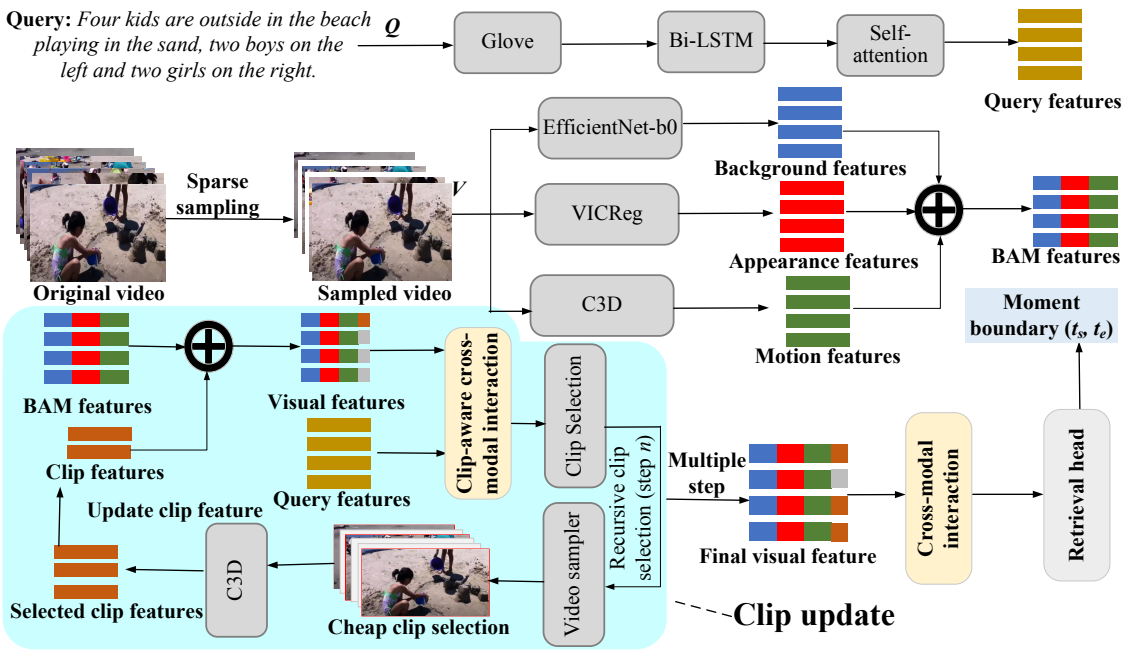

Figure 2: Overview of our model, which introduce a novel clip selection approach to search the core clip for VMR iteratively.

and Hu 2022) can be divided into two categories: 1) Top-down methods (Gao et al. 2017; Liu et al. 2022d; Zhang et al. 2020b; Zeng et al. 2020; Wang et al. 2022; Wang and Shi 2023; Wang, Jian, and Xue 2023; Wang et al. 2021b; Li et al. 2023a,b; Yang et al. 2023): They first pre-define multiple segment proposals and then align these proposals with the query for cross-modal semantic matching. Finally, the best proposal with the highest similarity score is selected as the predicted segment. Although achieving decent results, these proposal-based methods severely rely on the quality of the segment proposals and are time-consuming. 2) Bottom-up methods (Zhang et al. 2020a; Chen et al. 2020; Mun, Cho, and Han 2020; Tang et al. 2021; Nan et al. 2021; Ji et al. 2023a,b; Jian and Wang 2023): They directly regress the start and end boundary frames of the target segment or predict boundary probabilities frame-wisely. Compared with the proposal-based methods, proposal-free methods are more efficient. However, the above methods heavily rely on the datasets that require numerous manually labelled annotations for training. In real-world applications, we always collect overlong videos. If we directly utilize the video for VMR, it will lead to much computational cost. Although some egocentric video search works (Jia et al. 2022; Ramakrishnan, Al-Halah, and Grauman 2023) are proposed to accelerate the video understanding process, they rely on the egocentric video and a language question, which is different from the inputs of VMR. Thus, we present a brand-new setting, called clip trimming VMR, with a merely light-weighted model rather than a large-weighted network.

Proposed Method

We propose a light-weighted clip-selection approach that intelligently spots query-relevant clips for efficient VMR. Our model consists of a novel clip selection architecture called ClipSpotter, our BAM semantic indexing features to select latent query-relevant clips for cross-modal interaction, and distillation loss to address optimization issues arising from jointly training ClipSpotter with the VMR task modules. ClipSpotter previews the video using our RIO features, which are obtained by selecting a single image from each clip and encoding them using efficient visual encoders (Tan and Le 2019). Heavy clip features are then extracted from only the smaller subset of clips selected by ClipSpotter.

Overview

Problem statement. Given an untrimmed video $\mathcal{V}$ with the frame number of T and a sentence query $\mathcal{Q}$ composed of N words, the task of Video Moment Retrieval Using Language (VMR) aims to precisely locate a temporal moment boundary (τ_s, τ_e) in video $\mathcal{V}$, which starts at timestamp τ_s and ends at timestamp τ_e, according to the semantics of query $\mathcal{Q}$.

Pipeline. To tackle the light-weighted VMR, we propose a novel framework as shown in Figure 2. For convenience, we define continuous 16 frames as a clip and each clip overlaps 8 frames with adjacent clips. We denote the number of clips as $C = T/8 - 1$. First, a pretrained semantic indexer, which consists of one or more image encoders, is used to extract semantic index features. After encoding the visual and textual features, we feed the multi-modal features into a RetrievalSpotter module, which consists of a cross-modal interaction module and a selection policy to iteratively select the optimal clip. Especially, in each iteration, we update the clip feature as the final clip feature. After multiple steps, we can utilize the selective clip features for the VMR task.

BAM Features for Semantic Indexing

To adaptively preview the video and select query-relevant clips, we target to learn the semantic indexer based on three high-level visual features: background feature, appearance feature and motion feature, termed BAM features. An intuitive idea is to employ the efficient video recognition technology to extract the visual feature. However, previous efficient video recognition methods only capture the object-level appearance feature based on the pretrained ImageNet, and often ignore the background information and motion feature, which is insufficient for our challenging VMR task. Since query-aware indexing in the VMR requires the background feature and motion information, we try to integrate the query feature and three high-level features (background, appearance and motion features) for previewing video.

For example, in a query *"A woman and a man are sitting on the sidewalk playing music."*, the background-related text is *"sidewalk"*, the appearance-related text is *"a woman and a man"*, the motion-related text is *"playing music"*. All three kinds of visual information are significant for the VMR task. Hence, we design a set of low-cost semantic indexing features that capture context from the **b**ackground, object-level **a**ppearance, and **m**otion, named *BAM*.

Background features: To effectively capture the background characteristics, we utilize a pre-trained EfficientNet-b0 image encoder (Tan and Le 2019) as a background classifier. In the VMR task, a full video often contains less than three backgrounds, and the background change often corresponds to the start or end of an important video activity.

Appearance features: In a video, the view is often changed due to the object movement and the camera shift, which leads to the visual variance. To adaptively minimize the variance, we feed the frame-level video into a pretrained VICReg (Bardes, Ponce, and LeCun 2022) network in a self-supervised way. To extract appearance features accurately, we maintain diversity over each feature dimension, which can learn the object properties and closes the visual variance.

Motion features: For each clip, we first choose its start frame and end frame to extract the motion feature. Then, we feed the two frames into pretrained C3D network (Tran et al. 2015) to extract the clip-level motion features.

Overall, we sample one image within each video clip, extract each of the RIO features, and concatenate them to obtain the semantic indexing features $s = [s_1, s_2, \cdots, s_C] = SemanticIndexer(\mathcal{V})$. These are image features extracted by sampling one image within each video clip, and they are inexpensive to compute. These will serve as an initial preview of the video for intelligent clip selection.

Query features. Similarly, given the query Q, we also follow (Liu et al. 2022c, 2023e) to utilize the Glove (Pennington, Socher, and Manning 2014) embedding to encode each word into a dense vector. We further employ the Bi-GRU (Chung et al. 2014) layers to encode the word-level sequential information in the whole sentence. The final word-level feature can be denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_j\}_{j=1}^{N} \in \mathbb{R}^{N\times D}$.

$$q = [q_1, q_2, \cdots, q_N] = QueryEncoder(\mathcal{Q}). \tag{1}$$

Thus, by the semantic index s, query features q, and the video $\mathcal{V}$, we design a RetrievalSpotter module to recursively retrieve a final subset of query-relevant video clips $\mathcal{V}'$ as the moment candidates. Then, we feed all the frames of the selected clips $\mathcal{V}'$ into the same pre-trained C3D network to extract the expensive clip features. To avoid the unimportant computational cost, we set the features of those selected clips to zero in $v' = VideoEncoder(\mathcal{V}') \in \mathbb{R}^{C\times D}$.

RetrievalSpotter Architecture

To adaptively select the query-relevant clips in the VMR task, we design a RetrievalSpotter network, which first previews the entire video by an efficient semantic index, and then alternates between selected clips for expensive feature extraction. Finally, the RetrievalSpotter network repeats the above two processes until all the clips will be selected in the next recursive clip or the current recursive step reaches the maximum step. Thus, the semantic index s is computed once before step 1 and kept fixed.

Specifically, we denote the recursive step as $b \in [1, \cdots, B]$, where B is the maximum step. $v'_b \in \mathbb{R}^{C\times D}$ denote the clip features for selected from steps 1 to $b-1$, where $v'_1 = [0]_{C\times D}$ is an all-zero matrix. For any clip feature c_b, we concatenate it with s along the feature dimension to fuse them as $s_b \in \mathbb{R}^{C\times 2D}$. To interact the visual and textual features for further reasoning, we perform the clip-aware cross-modal fusion as: $c'_b = CrossModalFusion(s_b, q)$, where c'_b is the fused cross-modal feature.

By treating the fused feature as guidance, we utilize a two-layered MLP as the clip selection module to show if a clip feature should be computed or not. Especially, the clip selection module will output a binary value (*i.e.*, thresholded probabilities) for each clip:

$$p_{b+1} = ClipSelection(c'_b) \in \{0, 1\}^C, \tag{2}$$

where p_{b+1} is the binary value, where if $p_{b+1} = 1$, the clip will be selected and vice versa. Finally, we feed these selected clips into the video encoder (C3D) to obtain the expensive visual clip features. We repeat the selection process for steps, and utilize a cumulative set of clip features ($f_v = v'_{B+1}$) to predict the moment boundary.

To enhance the visual features with query-specific information, we perform cross-modal interaction by the concatenated clip and semantic index features $s \oplus v$ and the query features q.

$$c = CrossModalInteraction(s \oplus v, q) \in \mathbb{R}^{C\times D_h}. \tag{3}$$

Finally, we design a retrieval module to predict the temporal extent of the boundary $\hat{\mathcal{B}}$: $\hat{\mathcal{B}} = [\tau_s, \tau_e] = Retrieval(c)$.

Thanks to the above processes, we can enhance the efficiency of the state-of-the-art VMR methods. Specifically, the VMR works (*e.g.*, VSLNet (Zhang et al. 2020a) and MMN) feed the entire video into the visual encoder to extract the clip features, utilize a query encoder to extract query features, use a cross-model interaction module to fuse visual and textual feature, and finally employ a retrieval head module to predict the moment boundary. Our proposed SpotVMR can significantly save the computational cost by first previewing the video cheaply using the semantic indexer and then recursively selecting a subset of query-relevant clips by RetrievalSpotter. By iteratively selecting a subset of clips for expensive feature extraction,

our SpotVMR can modulate the video inputs to these state-of-the-art VMR models. Although different models utilize various cross-modal interaction modules and retrieval head modules, our efficient clip selection strategy still works.

Model Optimization

After selecting the query-relevant clips, we jointly optimize the cross-modal interaction and the retrieval head modules end-to-end to improve the retrieval performance. During training, we keep the video, semantic index, and text encoders frozen. Besides, multiple loss functions are introduced: a VMR task loss $\mathcal{L}_{vmr}$, a clip selection loss $\mathcal{L}_{sel}$, and a novel feature distillation loss $\mathcal{L}_{ftd}$.

Clip selection loss. To select accurate clips to avoid under-sampling and over-sampling, we introduce a clip selection loss: $L_{sel} = (\gamma - \mathbb{E}_{(v,q)\sim D_t}[\frac{1}{L}\sum_{l=1}^{L} b^l_{joint}])^2$, where D_t denotes the training dataset and $\boldsymbol{b}_{joint} = \sum_{n=1}^{N+1} p_b$ is the overall binary selections after N steps. By predefining the hyperparameter γ, the clip selection loss can limit the fraction of selected clips in expectation. To regularize the per-step clip selection p^l_b, we encourage our proposed model to select $(\gamma L/B)$ clips in each step. Experimental results show that the above simple regularization can significantly improve training stability. Since our proposed RetrievalSpotter predicts binary values during clip selection, it is not differentiable for gradient-based optimization. Therefore, we introduce the Gumbel-Softmax trick to reparameterize argmax sampling using a softmax relaxation during training (Hazan and Jaakkola 2012; Wu et al. 2019).

VMR task losses. We denote the video and query features as v and q. By fusing v and q, we utilize the cross-modal interaction module to obtain the cross-modal representation as $c = CrossModalInteraction(v, q) \in \mathbb{R}^{C\times D}$. Especially, the module includes a transformer encoder module, which utilizes the self-attention process to update the video feature v and query feature q independently. Then, we utilize the context-query attention mechanism for enhancing the video features with the help of the query features (Zhang et al. 2020a; Seo et al. 2016). Then, we introduce a 1D convolutional layer to compute the probability that a clip lies within a temporal neighborhood of the target moment: $\hat{\mathcal{S}}_h = \sigma(Conv1D(c)) \in \mathbb{R}^{C\times 1}$, where σ is the sigmoid function, and $\hat{\mathcal{S}}_h$ is used to update the cross-modal features: $c = \hat{\mathcal{S}}_h \cdot c \in \mathbb{R}^{C\times D_h}$. To infer the moment boundary, we introduce a retrieval module, including a transformer encoder for performing self-attention and an MLP layer to predict the log probabilities:

$$\hat{\tau}_s, \hat{\tau}_e = RetrievalPrediction(c), \tag{4}$$

where $\hat{\tau}_s, \hat{\tau}_e \in \mathbb{R}^{C\times 1}$ are log-probabilities per feature location, "RetrievalPrediction" means the retrieval module. We use the following loss to supervise the boundary predictions:

$$\mathcal{L}_{boundary} = L_{CE}(\hat{p}_s, p^*_s) + L_{CE}(\hat{p}_e, p^*_e), \tag{5}$$

where L_{CE} denotes the cross-entropy loss, and p^*_s, p^*_e are the ground-truth boundary of the target moment. We supervise the query-aware visual enhancement by the following loss:

$$\mathcal{L}_{qav} = f_{CE}(\hat{S}_h, S^*_h), \tag{6}$$

where S^*_h denotes the ground-truth enhancement score, which covers an extended temporal window around the ground-truth moment boundary. By jointing the above loss, we can obtain the overall VMR loss as follows:

$$\mathcal{L}_{vmr} = \mathcal{L}_{boundary} + \mathcal{L}_{qav}. \tag{7}$$

Distillation loss: To further fine-tune the moment boundary, we design a two-stage training strategy based on the knowledge distillation approach. First, we train a teacher VMR model without the RetrievalSpotter module as the distillation supervision. Then, we utilize a student VMR module with the RetrievalSpotter module for joint optimization.

Given a video-query pair (V, Q) and its ground-truth moment boundary B, we denote the cross-modal interaction outputs for the teacher and student VMR modules as $c_{teacher}$ and $c_{student}$, respectively. Different from the student module, we feed all the video features into the teacher module. To match the cross-modal features between the student module and the teacher module, we design the following feature distillation loss: $\mathcal{L}_{ftd} = ||StopGrad(c_{bteacher}) - c_{bstudent}||_2$, where $||\cdot||_2$ is the L-2 loss, the gradient is not propagated to the frozen teacher.

Thus, our final loss is:

$$\mathcal{L}_{final} = \alpha\mathcal{L}_{vmr} + \beta\mathcal{L}_{sel} + \gamma\mathcal{L}_{ftd}, \tag{8}$$

where α, β and γ are hyperparameters to balance the weights of different losses. By jointly training these losses, we encourage the model to improve VMR performance while limiting the budget of clips selected.

Experiments

Dataset. For fair comparison with existing VMR works, we utilize the same datasets for evaluation: ActivityNet Caption (Caba Heilbron et al. 2015), TACoS (Regneri et al. 2013), and Charades-STA (Sigurdsson et al. 2016). Specifically, ActivityNet Caption contains 20000 untrimmed videos with 100000 descriptions from YouTube. Following the public split, we use 37417, 17505, and 17031 sentence-video pairs for training, validation, and testing. TACoS contains 127 videos collected from cooking scenarios. We also follow the public split, which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing. As for Charades-STA, there are 12408 and 3720 moment-query pairs in the training and testing sets, respectively.

Evaluation metrics. Following (Gao et al. 2017; Zhang et al. 2020a), we adopt "R@n, IoU=m" as the evaluation metrics. The "R@n, IoU=m" denotes the percentage of language queries having at least one result whose Intersection over Union (IoU) with ground truth is larger than m in top-n retrieved segment. In our experiments, we use $n \in \{1, 5\}$ for all datasets, $m \in \{0.5, 0.7\}$ for ActivityNet Captions and Charades-STA, $m \in \{0.3, 0.5\}$ for TACoS.

Implementation details. To encode each video, we define continuous 16 frames as a clip and each clip overlaps 8 frames with adjacent clips. Following previous works (Zhang et al. 2020b; Wang et al. 2022), we employ the Glove model (Pennington, Socher, and Manning 2014) to embed each word to 300 dimension features. We train our whole

Performance comparisons on ActivityNet Captions					
Method	Type	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CTRL	↓	29.01	10.34	59.17	37.54
SCDM	↓	36.75	19.86	64.99	41.53
CMIN	↓	43.40	23.88	67.95	50.73
2D-TAN	↓	44.51	26.54	77.13	61.96
DRN	↓	45.45	24.36	77.97	50.30
MMN	↓	48.59	29.26	79.50	64.76
GDP	↑	39.27	-	-	-
LGI	↑	41.51	23.07	-	-
VSLNet	↑	43.22	26.16	-	-
IVG-DCL	↑	43.84	27.10	-	-
Ours	↕	**52.83**	**32.76**	**84.37**	**68.95**
Performance comparisons on Charades-STA					
Method	Type	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CTRL	↓	23.63	8.89	58.92	29.57
SCDM	↓	54.44	33.43	74.43	58.08
2D-TAN	↓	39.81	23.25	79.33	51.15
DRN	↓	53.09	31.75	89.06	60.05
MMN	↓	47.31	27.28	83.74	58.41
GDP	↑	39.47	18.49	-	-
VSLNet	↑	47.31	30.19	-	-
IVG-DCL	↑	50.24	32.88	-	-
ACRM	↑	57.53	38.33	-	-
Ours	↕	**68.82**	**47.39**	**97.01**	**75.38**
Performance comparisons on TACoS					
Method	Type	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
CTRL	↓	18.32	13.30	36.69	25.42
SCDM	↓	26.11	21.17	40.16	32.18
CMIN	↓	24.64	18.05	38.46	27.02
2D-TAN	↓	37.29	25.32	57.81	45.03
DRN	↓	-	23.17	-	33.36
MMN	↓	39.24	26.17	62.03	47.39
GDP	↑	24.14	-	-	-
VSLNet	↑	29.61	24.27	-	-
IVG-DCL	↑	38.84	29.07	-	-
ACRM	↑	38.79	26.94	-	-
Ours	↕	**48.72**	**38.94**	**67.03**	**56.38**

Table 1: Performance comparisons on three challenging datasets (top: ActivityNet Captions, middle: Charades-STA, bottom: TACoS), where ↓ means the top-down setting; ↑ means the bottom-up setting, and ↕ means our posed setting.

model for 100 epochs with an early stopping strategy. Parameter optimization is performed by Adam optimizer with a learning rate of 0.0005, and a linear decay rate of 1.0. All the experiments are implemented by PyTorch. For the hyper-parameters, we set $\alpha = 0.4$, $\beta = 0.8$, and $\gamma = 0.6$.

Comparison with State-of-the-Arts

For performance evaluation, we compare several state-of-the-art open-source VMR methods that are grouped into two categories: 1) Top-down (↓): CTRL (Gao et al. 2017), SCDM (Yuan et al. 2019), CMIN (Zhang et al. 2019b), 2D-TAN (Zhang et al. 2020b), DRN (Zeng et al. 2020), MMN (Wang et al. 2022); 2) Bottom-up (↑): GDP (Chen et al. 2020), LGI (Mun, Cho, and Han 2020), VSLNet (Zhang et al. 2020a), IVG-DCL (Nan et al. 2021), ACRM (Tang et al. 2021). The best results are **bold**. As shown in Table 1, our model beats all compared methods by a large margin, which illustrates the effectiveness of our model.

Model	T_{ext}				T_{exe}	T_{total}
	B	A	M	Other		
CTRL	-	-	-	18.51	187.52	206.03
RaNet	-	-	-	18.51	208.40	226.91
2D-TAN	-	-	-	18.51	216.87	235.38
MIGCN	-	-	-	18.51	253.94	271.89
MMN	-	-	-	18.51	289.31	207.82
DRN	-	-	-	18.51	294.70	313.21
Ours	**1.02**	**1.34**	**2.57**	-	**32.75**	**37.68**

Table 2: Efficiency comparison (time complexity (s) of 100 videos) on ActivityNet Captions. The total time T_{total} comprises the measurement time of extracting the corresponding features (T_{ext}), and executing the network models (T_{exe}), where "Other" means the feature encoder (*e.g.,* C3D/I3D).

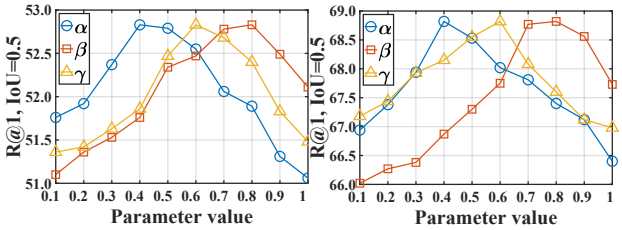

Figure 3: Analysis on the parameters (α, β, γ) on ActivityNet Captions (left) and Charades-STA (right).

Efficiency comparison. As shown in Table 2, we conduct the efficiency comparison on ActivityNet Captions with some state-of-the-art open-source methods. Our model is more efficient than compared methods by a large margin.

Ablation Study

Main ablation study. To demonstrate the effectiveness of each component in our model, we conduct ablation studies regarding the components. The corresponding experimental results are reported in Table 3. Obviously, we can find that both two modules contribute a lot to the final performances, showing that each module is effective for the VMR task.

Plug-and-play. To further compare with current methods, we serve our method as a plug-and-play module for state-of-the-art models (2D-TAN and MMN). As shown in Table 4, our method can significantly improve their performance, which shows the effectiveness of our method.

Effect of the BAM feature. To analyze the contribution of different high-level features, we conduct the ablation study in Table 5. Background(B), appearance(A) and motion(M) features can significantly improve the performance. The improvement shows the effectiveness of our designed features. During the clip selection, we set the maximum recursive step as $B_{\max}$, we analyze the effect of different $B_{\max}$ on Table 6. When $B_{\max} = 5$, we can obtain the best performance.

Analysis on parameters. We conduct the ablation stud-

Method	ActivityNet				Charades				TACoS			
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
w/o $\mathcal{L}_{qav}$	48.72	29.53	82.53	67.30	68.01	45.32	94.27	73.52	46.80	37.85	65.29	53.71
w/o $\mathcal{L}_{sel}$	50.66	30.80	81.32	66.97	67.35	44.86	95.41	74.25	47.13	37.92	66.17	54.82
w/o $\mathcal{L}_{ftd}$	51.39	31.95	84.03	68.54	67.90	46.81	96.57	75.32	48.08	38.75	66.88	54.18
Full model	**52.83**	**32.76**	**84.37**	**68.95**	**68.82**	**47.39**	**97.01**	**75.38**	**48.72**	**38.94**	**67.03**	**56.38**

Table 3: Main ablation study on all the datasets, where we remove each key individual module to investigate its contribution.

Model	Setting	ActivityNet Captions				Charades-STA				TACoS			
		R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
2D-TAN	Origin	44.51	26.54	77.13	61.96	39.81	23.25	79.33	51.15	37.29	25.32	57.81	45.03
	Ours	**46.21**	**27.43**	**78.64**	**63.58**	**42.95**	**25.10**	**81.72**	**53.27**	**38.49**	**26.15**	**59.22**	**45.83**
MMN	Origin	48.59	29.26	79.50	64.76	47.31	27.28	83.74	58.41	39.24	26.17	62.03	47.39
	Ours	**49.37**	**30.52**	**80.26**	**65.31**	**49.12**	**28.54**	**85.43**	**59.72**	**41.17**	**27.43**	**63.92**	**49.30**

Table 4: Our proposed method serves as a plug-and-play module for state-of-the-art models on different datasets.

B	A	M	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
✔	✗	✔	50.24	30.62	82.49	68.43
✔	✔	✗	51.35	31.84	82.88	68.27
✗	✔	✔	51.98	32.25	83.71	68.60
✔	✔	✔	**52.83**	**32.76**	**84.37**	**68.95**

Table 5: Effect of semantic index on ActivityNet Captions.

Module	Changes	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Recursive steps	3	51.92	31.40	83.55	67.38
	5	52.83	**32.76**	**84.37**	**68.95**
	7	**53.04**	31.95	83.52	68.24

Table 6: Effect of recursive step on ActivityNet Captions.

ies on the parameters α, β, γ in Figure 3. Specifically, we change one parameter with fixing the others. We obtain the best performance when $\alpha = 0.4, \beta = 0.6, \gamma = 0.8$.

Qualitative results

We provide the retrieval visualizations on three datasets in Figure 4. Our method can retrieve more precise moment boundaries than previous state-of-the-art methods (MMN (Wang et al. 2022) and WSTAN (Wang et al. 2021a)).

Conclusion

In this paper, we propose a novel and efficient video moment retrieval setting, which first previews the whole video by a semantic indexer, and then retrieves the target moment boundary by a distillation loss. Experiments on three challenging datasets show the effectiveness of our method.

Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC) under grant no. 61972448 and no. 62272328.

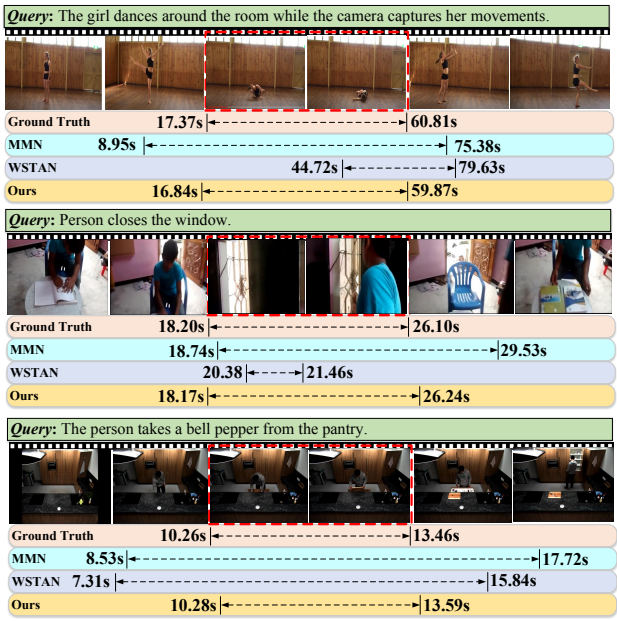

Figure 4: Qualitative results on three datasets (top: ActivityNet Captions, middle: Charades-STA, bottom: TACoS).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*.

Bardes, A.; Ponce, J.; and LeCun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In *ICLR*.

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

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *EMNLP*.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li,

X. 2020. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *AAAI*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS*.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially Relevant Video Retrieval. In *MM*.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2022b. Dual encoding for video retrieval by text. *TPAMI*.

Dong, J.; Peng, X.; Ma, Z.; Liu, D.; Qu, X.; Yang, X.; Zhu, J.; and Liu, B. 2023a. From Region to Patch: Attribute-Aware Foreground-Background Contrastive Learning for Fine-Grained Fashion Retrieval. *arXiv*.

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022c. Reading-strategy inspired visual representation learning for text-to-video retrieval. *TCSVT*.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023b. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval. In *CVPR*.

Fang, X.; and Hu, Y. 2020. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. 2021a. ANIMC: A Soft Approach for Autoweighted Noisy and Incomplete Multiview Clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2020. V^3H: View variation and view heredity for incomplete multiview clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *TETCI*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-Modal Cross-Domain Alignment Network for Video Moment Retrieval. *TMM*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *CVPR*, 2448–2460.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023c. Hierarchical local-global transformer for temporal sentence grounding. *TMM*.

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

Ge, R.; Gao, J.; Chen, K.; and Nevatia, R. 2019. Mac: Mining activity concepts for language-based temporal localization. In *WACV*.

Hazan, T.; and Jaakkola, T. S. 2012. On the Partition Function and Random Maximum A-Posteriori Perturbations. In *ICML*.

Ji, C.; Li, J.; Peng, H.; Wu, J.; Fu, X.; Sun, Q.; and Yu, P. S. 2023a. Unbiased and Efficient Self-Supervised Incremental Contrastive Learning. In *WSDM*.

Ji, C.; Zhao, T.; Sun, Q.; Fu, X.; and Li, J. 2023b. Higher-Order Memory Guided Temporal Random Walk for Dynamic Heterogeneous Network Embedding. *PR*.

Ji, W.; Chen, L.; Wei, Y.; Wu, Y.; and Chua, T.-S. 2023c. Mrtnet: Multi-resolution temporal network for video sentence grounding. *ICASSP*.

Ji, W.; Li, L.; Fei, H.; Liu, X.; Yang, X.; Li, J.; and Zimmermann, R. 2023d. Towards Complex-query Referring Image Segmentation: A Novel Benchmark. *arXiv*.

Ji, W.; Li, X.; Wei, L.; Wu, F.; and Zhuang, Y. 2020. Context-aware graph label propagation network for saliency detection. *TIP*.

Ji, W.; Li, X.; Wu, F.; Pan, Z.; and Zhuang, Y. 2019. Human-centric clothing segmentation via deformable semantic locality-preserving network. *TCSVT*.

Ji, W.; Li, X.; Zhuang, Y.; Bourahla, O. E. F.; Ji, Y.; Li, S.; and Cui, J. 2018. Semantic Locality-Aware Deformable Network for Clothing Segmentation. In *IJCAI*.

Ji, W.; Li, Y.; Wei, M.; Shang, X.; Xiao, J.; Ren, T.; and Chua, T.-S. 2021. Vidvrd 2021: The third grand challenge on video relation detection. In *MM*.

Ji, W.; Liang, R.; Liao, L.; Fei, H.; and Feng, F. 2023e. Partial annotation-based video moment retrieval via iterative learning. In *MM*.

Ji, W.; Liang, R.; Zheng, Z.; Zhang, W.; Zhang, S.; Li, J.; Li, M.; and Chua, T.-s. 2023f. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *CVPR*.

Ji, W.; Liu, X.; Zhang, A.; Wei, Y.; and Wang, X. 2023g. Online Distillation-enhanced Multi-modal Transformer for Sequential Recommendation. In *MM*.

Jia, B.; Lei, T.; Zhu, S.-C.; and Huang, S. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*.

Jian, X.; and Wang, Y. 2023. InvGC: Robust Cross-Modal Retrieval by Inverse Graph Convolution. In *Findings of EMNLP*.

Li, Q.; Guo, S.; Ji, C.; Peng, X.; Cui, S.; and Li, J. 2023a. Dual-Gated Fusion with Prefix-Tuning for Multi-Modal Relation Extraction. *Finding of ACL*.

Li, Q.; Guo, S.; Luo, Y.; Ji, C.; Wang, L.; Sheng, J.; and Li, J. 2023b. Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment. *WWW*.

Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023a. DICNet: Deep Instance-Level Contrastive Network for Double Incomplete Multi-View Multi-Label Classification. In *AAAI*.

Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023b. Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers. In *AAAI*.

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023c. Information Recovery-Driven Deep Incomplete Multiview Clustering Network. *TNNLS*.

Liu, D.; Fang, X.; Hu, W.; and Zhou, P. 2023d. Exploring Optical-Flow-Guided Motion and Detection-Based Appearance for Temporal Sentence Grounding. *TMM*.

Liu, D.; Fang, X.; Zhou, P.; Di, X.; Lu, W.; and Cheng, Y. 2023e. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*.

Liu, D.; and Hu, W. 2022. Skimming, Locating, then Perusing: A Human-Like Framework for Natural Language Video Localization. In *MM*.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*.

Liu, D.; Qu, X.; Dong, J.; Nan, G.; Zhou, P.; Xu, Z.; Chen, L.; Yan, H.; and Cheng, Y. 2023f. Filling the Information Gap between Video and Query for Language-Driven Moment Retrieval. In *MM*.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2021a. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *EMNLP*.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021b. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *CVPR*.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Xu, Z.; Wang, H.; Di, X.; Lu, W.; and Cheng, Y. 2023g. Transform-Equivariant Consistency Learning for Temporal Sentence Grounding. *TOMM*.

Liu, D.; Qu, X.; and Hu, W. 2022. Reducing the Vision and Language Bias for Temporal Sentence Grounding. In *MM*.

Liu, D.; Qu, X.; Liu, X.-Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *MM*.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022b. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*.

Liu, D.; Qu, X.; and Zhou, P. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding. In *EMNLP*.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022c. Exploring Motion and Appearance Information for Temporal Sentence Grounding. In *AAAI*.

Liu, D.; Zhou, P.; Xu, Z.; Wang, H.; and Li, R. 2022d. Few-Shot Temporal Sentence Grounding via Memory-Guided Semantic Learning. *TCSVT*.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *MM*.

Liu, Y.; Wang, K.; Shao, W.; Luo, P.; Qiao, Y.; Shou, M. Z.; Zhang, K.; and You, Y. 2023h. MLLMs-Augmented Visual-Language Representation Learning. *arXiv*.

Ma, Z.; Dong, J.; Long, Z.; Zhang, Y.; He, Y.; Xue, H.; and Ji, S. 2020. Fine-grained fashion similarity learning by attribute-specific embedding network. In *AAAI*.

Mun, J.; Cho, M.; and Han, B. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *CVPR*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Qu, X.; Tang, P.; Zou, Z.; Cheng, Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *MM*.

Qu, X.; Zeng, J.; Liu, D.; Wang, Z.; Huai, B.; and Zhou, P. 2023. Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In *AAAI*.

Qu, X.; Zou, Z.; Su, X.; Zhou, P.; Wei, W.; Wen, S.; and Wu, D. 2021. Attend to where and when: Cascaded attention network for facial expression recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(3): 580–592.

Ramakrishnan, S. K.; Al-Halah, Z.; and Grauman, K. 2023. SpotEM: efficient video search for episodic memory. In *ICML*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL*.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv*.

Shen, X.; Zhang, X.; Yang, X.; Zhan, Y.; Lan, L.; Dong, J.; and Wu, H. 2023. Semantics-Enriched Cross-Modal Alignment for Complex-Query Video Moment Retrieval. In *MM*.

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

Sun, S.; Liu, D.; Dong, J.; Qu, X.; Gao, J.; Yang, X.; Wang, X.; and Wang, M. 2023. Unified Multi-modal Unsupervised Representation Learning for Skeleton-based Action Understanding. In *MM*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*.

Tang, H.; Zhu, J.; Liu, M.; Gao, Z.; and Cheng, Z. 2021. Frame-wise Cross-modal Matching for Video Moment Retrieval. *TMM*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Wang, Y.; Deng, J.; Zhou, W.; and Li, H. 2021a. Weakly supervised temporal adjacent network for language grounding. *TMM*.

Wang, Y.; Jian, X.; and Xue, B. 2023. Balance Act: Mitigating Hubness in Cross-Modal Retrieval with Query and Gallery Banks. In *EMNLP*.

Wang, Y.; and Shi, P. 2023. Video-Text Retrieval by Supervised Sparse Multi-Grained Learning. In *Findings EMNLP*.

Wang, Y.; Xue, B.; Cheng, Q.; Chen, Y.; and Zhang, L. 2021b. Deep Unified Cross-Modality Hashing by Pairwise Data Alignment. In *IJCAI*.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023a. Deep Double Incomplete Multi-View Multi-Label Learning With Incomplete Labels and Missing Views. *TNNLS*.

Wen, J.; Yan, K.; Zhang, Z.; Xu, Y.; Wang, J.; Fei, L.; and Zhang, B. 2021. Adaptive graph completion based incomplete multi-view clustering. *TMM*.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2023b. A survey on incomplete multiview clustering. *TSYST MAN CY-S*.

Wu, Z.; Xiong, C.; Jiang, Y.-G.; and Davis, L. S. 2019. LiteEval: A Coarse-to-Fine Framework for Resource Efficient Video Recognition. In *NeurIPS*.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *CVPR*.

Yang, X.; Wang, S.; Dong, J.; Dong, J.; Wang, M.; and Chua, T.-S. 2022. Video moment retrieval with cross-modal neural architecture search. *TIP*.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NeurIPS*.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *CVPR*.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based Localizing Network for Natural Language Video Localization. In *ACL*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *TOMM*.