This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fast Video Moment Retrieval

Junyu Gao[1,2] and Changsheng Xu[1,2,3]
[1] National Lab of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Peng Cheng Laboratory, ShenZhen, China
{junyu.gao, csxu}@nlpr.ia.ac.cn

Abstract

This paper targets at fast video moment retrieval (fast VMR), aiming to localize the target moment efficiently and accurately as queried by a given natural language sentence. We argue that most existing VMR approaches can be divided into three modules namely video encoder, text encoder, and cross-modal interaction module, where the last module is the test-time computational bottleneck. To tackle this issue, we replace the cross-modal interaction module with a cross-modal common space, in which moment-query alignment is learned and efficient moment search can be performed. For the sake of robustness in the learned space, we propose a fine-grained semantic distillation framework to transfer knowledge from additional semantic structures. Specifically, we build a semantic role tree that decomposes a query sentence into different phrases (subtrees). A hierarchical semantic-guided attention module is designed to perform message propagation across the whole tree and yield discriminative features. Finally, the important and discriminative semantics are transferred to the common space by a matching-score distillation process. Extensive experimental results on three popular VMR benchmarks demonstrate that our proposed method enjoys the merits of high speed and significant performance.

1. Introduction

Video Moment Retrieval (VMR) aims to localize a temporal segment from an untrimmed video, as queried by a natural language sentence [14, 1]. It plays a crucial role in video understanding and has various downstream applications such as robotic navigation, autonomous driving, video entertainment, and so forth. Despite great successes in recent years [63, 57, 65, 40, 62, 64, 38], effective VMR remains challenging due to many factors including complex video scenes, fine-grained semantic query structures, and huge cross-modal gap between visual and textual features.

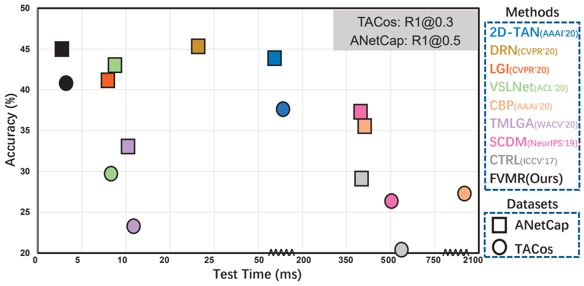


Figure 1. Test time and accuracy plot of state-of-the-art VMR approaches on TACos and ActivityNet Captions (ANetCap). We report the metrics R1@0.3 and R1@0.5 for the two datasets respectively. Our proposed FVMR achieves the best accuracy-speed balance among all the competitors. Best viewed in color.

To tackle the above challenges, the current state-of-the-art VMR pipeline can be divided into three modules namely video encoder, text encoder, and cross-modal interaction module. The first two encoders utilize convolutional neural networks (*e.g*. C3D [49] or I3D [4]) and recurrent neural networks (*e.g*. BiLSTM [27] or GRU [12]) to extract visual and textual features respectively. Then, to predict the target moment, the cross-modal interaction module is designed to jointly consider both modalities by using different architectures such as cross attention [36, 35, 7, 64], graph neural networks [63, 67, 42], and temporal adjacent networks [65].

Despite the above achievements, we emphasize, fast video moment retrieval (*fast VMR*) is in fact often necessary, since localizing the target moment is usually employed only as a part of time-critical video retrieval systems. For example, to localize moments in a video corpus [46, 30], it usually requires us to perform efficient VMR on hundreds to thousands of candidate videos for a given natural language query. Also, fast video moment retrieval on embedded devices may enable many additional applications, such as intelligent robot service and smart home [68]. However, as shown in Figure 1, high-speed and effective VMR algorithms remain scarce. Taking the recently state-of-the-art approach 2D-TAN [65] as an example, it will cost about 100 milliseconds when performing VMR on one single video.

We argue that among the three modules in the VMR pipeline, cross-modal interaction is the test-time computational bottleneck. The reasons are three-fold: **(1)** Before model testing, it is convenient to pre-extract and store video features in an offline manner. As a result, the video encoder does not affect the test time. **(2)** The text encoder is highly efficient ($\sim$3 ms per sentence) and irreplaceable for all the VMR methods. Therefore, targeting at a low cost of text encoder is unadvisable. **(3)** Cross-modal interaction takes up most of test time due to the complex feature fusion operation [54, 61, 34, 62, 63, 40, 42, 8] and subsequent feature transformations [65, 50, 23]. The above observations motivate us to design an efficient and effective cross-modal interaction module. Ideally, this module can be simplified to a cross-modal common space, where moment-query alignment is learned. In such a space, for the given query, it is nearly cost-free to obtain the matching score of each moment proposal, *i.e.*, we can simply calculate the similarities between video moment features and the query sentence feature in the common space by using efficient vector operations like dot product. A few early work [1, 26] has explored common space learning for VMR. However, their performance is far below the current state-of-the-arts. In fact, without a well-designed cross-modal interaction, it is difficult to ground a textual query onto the video effectively. Therefore, the common space learning strategy has become an underdog in VMR, which motivates us to solve the following problem: *How to learn a common space that can not only yield efficient moment/query features for fast VMR but also improve the discriminative ability by leveraging fine-grained semantic structures?*

To this end, we propose a fine-grained semantic distillation framework for fast video moment retrieval, which learns an efficient and effective moment-query common space by transferring knowledge from additional semantic structures. Specifically, our proposed approach consists of four modules, namely video encoder, text encoder, fine-grained semantic extractor, and common space. Here, in addition to the text encoder (a Bi-LSTM in our proposed approach), we introduce a fine-grained semantic extractor to facilitate the learning of common space. This extractor decomposes a query sentence into fine-grained semantic structures (phrases) by building a semantic role tree, where each phrase is represented as a subtree. Then, a hierarchical semantic-guided attention module is designed to propagate semantic information across the whole tree and yield discriminative features for each phrase. Note that the learned fine-grained phrase features serve as complementary cues to provide an enhanced supervisory signal when learning the common space. During model training, the video and text encoders are required to learn from the fine-grained semantic extractor by matching score distillation. As a result, fine-grained semantic information is injected into the common space for robust moment-query alignment. During testing, we only exploit the text and video encoders to perform VMR, which does not adds computational overhead. As shown in Figure 1, our proposed method achieves the best accuracy-speed balance among state-of-the-arts.

The main contributions of this paper are three-fold:

- We introduce fast video moment retrieval (FVMR) that aims for retrieving target moments efficiently and accurately. To this end, a simple yet effective common space learning paradigm is designed, not only speeding up VMR, but also improving the performance.
- We design a novel fine-grained semantic distillation framework for FVMR. Here, a hierarchical semantic-guided attention module is designed to leverage the fine-grained semantic structures by optimizing a matching-score distillation loss.
- Extensive experimental results on three popular VMR benchmarks demonstrate that our proposed method enjoys the merits of high speed and significant performance. Compared with the recent state-of-the-arts, 2D-TAN [65], our proposed model is $40\times$ faster and obtains 5.5% absolute gains on the TACos dataset [44].

2. Related Work

Video moment retrieval is to localize the correct moment in an untrimmed video that is semantically aligned with a given natural language query [14, 1, 59, 15]. It plays a crucial role in the video understanding field [13, 20, 19, 52, 17, 18, 21, 16]. Researchers have been proposing a variety of VMR approaches in either one-stage or two-stage manners. The one-stage methods [23, 61, 37, 40, 62, 45, 55, 24] aim to build a proposal-free framework and directly regress the temporal location of target moment by using the fused video and textual features. The Extractive Clip Localization (ExCL) method [23] directly uses recurrent networks to predict the start and end time by leveraging the cross-modal interaction between the text and video. Mun *et al.* design a Local-Global video-text Interaction algorithm (LGI) [40], which uses a sequential query attention module and exploits the implicit semantic information from local to global. Although one-stage approaches are efficient with favorable performance, most of them can only regress one temporal segment, which is not appropriate enough for the practical retrieval task. Different from the one-stage formulation, the current dominant approaches belong to the two-stage paradigm [54, 35, 9, 6, 3, 5, 60, 65, 50, 51, 2, 43, 34], which firstly generates moment proposals from the input video and then performs cross-modal fusion on each of the proposals to obtain the matching scores. Recent progress demonstrates that the two-stage strategy can not only generate diverse moment proposals but also achieve significant retrieval performance. Gao *et al*. propose the Cross-modal Temporal Regression Localizer (CTRL) [14] by using sliding windows to generate proposals. To improve the quality of the proposals, 2D-TAN [65] leverages a two-dimensional map to model the temporal relations between video moments with a temporal adjacent network. To improve the cross-modal interaction, Yuan *et al*. propose the Semantic Conditioned Dynamic Modulation (SCDM) [60] algorithm which modulates the temporally convolved visual features with the sentence semantics to correlate the sentence-related video contents. Other strategies are also adopted to improve the performance of video moment retrieval, such as graph neural networks [34, 63], reinforcement learning [2, 53, 25], weakly-supervised learning [39, 31], boundary-aware prediction [51], sentence reconstruction [32], and tree LSTM [66]. Until now, fast video moment retrieval with high performance is yet to be explored. Although early common space-learning algorithms [1, 26] and skip scanning-base method [24] can save computational cost during retrieval, they cannot get benefit from the explicitly fine-grained semantic information and the performance is much less than current state-of-the-arts. In this paper, we propose a simple yet effective method for fast video moment retrieval, not only speeding up VMR but also improving the performance.

Many existing VMR approaches only encode the semantic information of the query in a global manner [57, 25, 22, 14, 65, 53, 6], which cannot take full advantage of the intrinsic and fine-grained structure of the sentence. Although some approaches [22, 28, 10, 40, 43, 33, 56] exploit the semantic structure of a sentence, most of them only consider the semantics in a partial (*e.g*. activities or objects) [22, 28, 10] or an implicit manner [40, 43]. The CMIN method [67] leverages the syntactic structure of natural language queries by constructing a syntactic dependency graph. Chen *et al*. [11] propose to decompose a sentence as a semantic graph and integrate video-text matching at different levels including the global level, action level, and entity level. However, they [67, 11, 56] ignore to explicitly model the phrase-level structures. Moreover, the work [11] is designed for video-text retrieval, which is not suitable for VMR. For improving cross-modal interaction, [33] takes a modular network to model compositional natural language descriptions of activity in videos. Nevertheless, the complex interactions are not efficient. In this paper, targeting at fast video moment retrieval, we propose a fine-grained semantic distillation framework that explicitly leverages both global- and phrase-level structures.

3. Fast Video Moment Retrieval

Given a natural language query, this work aims to localize the target moment from an untrimmed video efficiently and accurately. To this end, we propose a fine-grained semantic distillation framework, which learns a moment-query common space by transferring knowledge from additional semantic structures. As shown in Figure 2, our proposed approach consists of four modules, namely video encoder, text encoder, fine-grained semantic extractor, and moment-query common space. In the following, we first present the video encoder and text encoder, then we introduce the fine-grained semantic extractor, which leverages complementary cues to provide an enhanced supervisory signal to the VMR task. In the fine-grained semantic extractor, we build a semantic role tree that decomposes a query sentence into different phrases (subtrees). A hierarchical semantic-guided attention module is designed to propagate semantic information across the whole tree and yield discriminative features for each phrase. Finally, we design a moment-query common space, where the video and text encoders are required to learn from the fine-grained semantic extractor by matching score distillation. During testing, the common space is adopted for fast video moment retrieval.

3.1. Video and Text Encoders

Video Encoder. For a given untrimmed video V, we firstly generate a set of moment proposals $P = \{p_i\}_{i=1}^N$, where p_i represents a proposal and N means the number of proposals. Then, the video encoder is utilized to extract the visual features of each moment proposal as follows:

$$\mathbf{M} = \{\mathbf{m}_1, ..., \mathbf{m}_i, ..., \mathbf{m}_N\} = \mathbf{Encoder}\left(\{p_i\}_{i=1}^N\right) \quad (1)$$

where $\mathbf{m}_i$ is the visual feature of proposal p_i, $\mathbf{m}_i \in \mathrm{R}^{D_v}$, and D_v is the dimension of the visual feature. In our framework, the video encoder can be any types of neural networks such as C3D [49] or I3D [4].

Text Encoder. For the query sentence $S = \{s_1, ..., s_L\}$ with L words, we simply employ a bidirectional LSTM (Bi-LSTM) [27] to obtain a sequence of word features $\{\mathbf{w}_1, ..., \mathbf{w}_L\}$ as follows:

$$\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_L = \mathbf{BiLSTM}(S), \quad (2)$$

where $\mathbf{w}_l = \overrightarrow{\mathbf{w}}_l \| \overleftarrow{\mathbf{w}}_l$ is the concatenation of the forward and backward hidden states of the BiLSTM for the l-th word. We jointly consider the addition of the beginning and the end features as the sentence feature, $\mathbf{s} = \mathbf{w}_1 + \mathbf{w}_L$, where $\mathbf{s} \in \mathrm{R}^{D_s}$.

3.2. Fine-grained Semantic Extractor

Numerous existing VMR methods [57, 25, 22, 14, 65, 53, 6] only adopt the extracted global sentence feature to perform temporal localization but ignore the intrinsic and fine-grained structure of the sentence. Obviously, as shown in Figure 2, a query sentence (*e.g*. "the man leaves the ring and the wrestler approaches the other wrestler and beats him to the ground") corresponding to a specific video moment has multiple semantic structures in global-level (the whole sentence) and phrase-level ("the man leaves the ring",

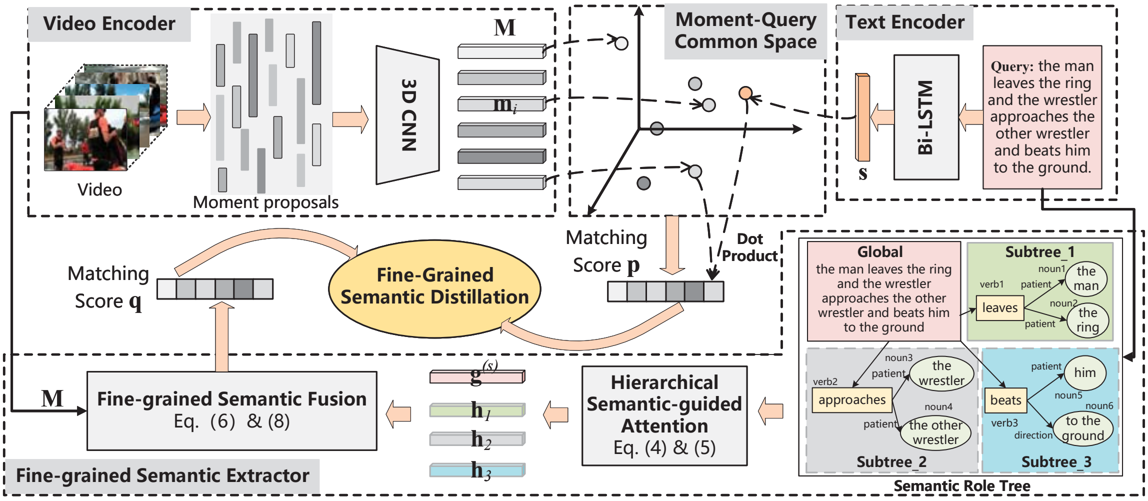


Figure 2. Overview of the proposed framework. The proposed fast video moment retrieval (FVMR) approach consists of four components including a video encoder, a text encoder, a fine-grained semantic extractor, and a moment-query common space. The two encoders extracts visual and textual features for moment proposals and the input query, which are then projected into the common space for the calculation of matching score $\mathbf{p}$. The fine-grained semantic extractor decomposes a query sentence into a semantic role tree, where each phrase is represented as a subtree. Then, a hierarchical semantic-guided attention module and a fine-grained semantic fusion module are designed to yield the other matching score $\mathbf{q}$. During model training, the video and text encoders are required to learn from the fine-grained semantic extractor by matching score distillation. During testing, only $\mathbf{p}$ is employed for fast video moment retrieval.

"the wrestler approaches the other wrestler" and "beats him to the ground"). The multiple semantic structures involve complicated interactions, which are actually organized as a semantic tree, indicating that a query sentence can be effectively grounded onto the video by properly aligning different semantic levels with the corresponding video parts. As a result, following [56], we build a semantic role tree to take full advantage of these details by using a semantic role labeling toolkit [47]. Semantic role labeling (SRL) comes from relation extraction, which aims to obtain predicates and arguments and determine how these arguments are semantically related to the predicate. Such semantic relations play an important role in comprehending the sentence.

For the query sentence, as shown in Figure 2, SRL parses it to predicates and arguments with semantic roles, where predicates are often verbs (*e.g.* actions) and arguments are often nouns (*e.g.* objects and entities). Then we set these verbs and nouns as nodes in our semantic role tree where the whole sentence is considered as the root node. All the verb nodes are connected with the root node, and noun nodes are set as leaf nodes. If a noun is related to a verb in the semantic aspect, we connect the two nodes. Similar to [11], we utilize a GRU [12] to obtain the initial node embeddings $\mathbf{g}^{(s)}$, $\mathbf{g}^{(v)}$ and $\mathbf{g}^{(n)}$, representing global embedding, predicate (verb) embeddings and argument (noun) embeddings respectively. Note that $\mathbf{g}^{(v)} = \{\mathbf{g}_1^{(v)}, ..., \mathbf{g}_{N^{(v)}}^{(v)}\}$ and $\mathbf{g}^{(n)} = \{\mathbf{g}_1^{(n)}, ..., \mathbf{g}_{N^{(n)}}^{(n)}\}$, $N^{(v)}$ and $N^{(n)}$ denote the number of verb and noun nodes, $\mathbf{g}^{(s)}, \mathbf{g}_i^{(v)}, \mathbf{g}_i^{(n)} \in \mathrm{R}^{D_f}$. The number of phrases is the same as the number of verbs.

Since a query sentence is composed of multiple phrases with fine-grained semantic structures, we aim to explicitly learn the discriminative phrase features for facilitating precise video moment retrieval. Note that the query is organized as a tree structure, where the top level provides guidance information for the lower level. As a result, we design a hierarchical semantic-guided attention module to leverage the intrinsic structure in an end-to-end manner. Specifically, features from the top level are adopted to estimate the importance scores of nodes in the lower level:

$$\begin{aligned} \boldsymbol{\alpha}_k^{(j)} &= \mathbf{W}_\alpha \left(\tanh(\mathbf{W}_{\text{top}}\hat{\mathbf{g}}^{(i)} \| \mathbf{W}_{\text{low}}\mathbf{g}_k^{(j)})\right), \\ \hat{\mathbf{g}}^{(i)} &= \sum_{l=1}^{N^{(i)}} \mathbf{a}_l^{(i)} \mathbf{g}_l^{(i)}, \end{aligned} \tag{3}$$

where $(i, j) \in \{(s, v), (v, n)\}$, indicating two types of consecutive hierarchy in the three-level semantic tree. $\mathbf{W}_\alpha \in \mathrm{R}^{1 \times 2D_f}$, $\mathbf{W}_{\text{top}} \in \mathrm{R}^{D_f \times D_f}$, and $\mathbf{W}_{\text{low}} \in \mathrm{R}^{D_f \times D_f}$ are learnable embedding matrices in the hierarchical semantic-guided attention module. $\tanh(\cdot)$ is the hyperbolic tangent activation function. $\mathbf{a}^{(i)} = \text{softmax}(\boldsymbol{\alpha}^{(i)})$. With the learned importance scores $\boldsymbol{\alpha}_k^{(j)}$, the feature of each phrase can be adaptively calculated in an attention manner:

$$\begin{aligned} \mathbf{h}_i &= \mathbf{b}_{i,1}\mathbf{g}_i^{(v)} + \sum_{j=2}^{\mathcal{N}_i+1} \mathbf{b}_{i,j}\mathbf{g}_{z_{i,j}}^{(n)}, i \in [1, ..., N_v], \\ \mathbf{b}_i &= \text{softmax}([\boldsymbol{\alpha}_i^{(v)}, \boldsymbol{\alpha}_{z_{i,1}}^{(n)}, ..., \boldsymbol{\alpha}_{z_{i,\mathcal{N}_i}}^{(n)}]), \end{aligned} \tag{4}$$

where $\mathcal{N}_i$ is the number of noun nodes connected with the

i-th verb node in the semantic role tree, and $z_{i,j}$ is the corresponding index of the noun node. Finally, we incorporate all the phrase features with the global embedding:

$$\mathbf{u} = \mathbf{g}^{(s)} \odot \frac{1}{N^{(v)}} \sum_{i=1}^{N^{(v)}} \mathbf{h}_i, \tag{5}$$

where $\odot$ is the Hadamard product operator, $\mathbf{u} \in \mathrm{R}^{D_u}$ is the learned fine-grained semantic feature. By using Eq. (5), both the global query information and the local phrase information are leveraged, which is exploited for the following common space learning.

3.3. Moment-Query Common Space Learning via Fine-Grained Semantic Distillation

Moment-Query Common Space. Our goal is to learn a moment-query common space, where video moment retrieval can be efficiently and effectively performed by vector similarity calculation. To this end, we adopt two feature transformation modules, ϕ_m and ϕ_s, to project the moment and query features into the common space. In order to perform fine-grained semantic distillation, we also project the fine-grained semantic feature into this space by using another feature transformation ϕ_u. In our framework, for simplicity, we utilize three multi-layer perceptrons (MLPs) for implementing ϕ_m, ϕ_s, ϕ_u. All the features in this space are D-dimensional. As a result, two types of matching scores can be calculated as follows:

$$\mathbf{p}_i = \phi_m(\mathbf{m}_i)^\top \phi_s(\mathbf{s}), \tag{6}$$

$$\mathbf{q}_i = \phi_{fuse}\left(\phi_m(\mathbf{m}_i) \odot \phi_u(\mathbf{u})\right), \tag{7}$$

where ϕ_{fuse} is an MLP. It learns the matching score $\mathbf{q}_i$ by using the fused moment and fine-grained semantic features. Note that we simply use the dot product to calculate $\mathbf{p}_i$ for fast moment retrieval, while we additionally adopt ϕ_{fuse} to further consider the interaction between moment and fine-grained semantic features. Since $\mathbf{q}_i$ exploits the fine-grained interaction, it is served as the teacher for the following fine-grained semantic distillation.

Video Moment Retrieval Loss. Because different moment proposals have different lengths, we compute the IoU score o_i for each proposal with the ground truth moment. Similar to [65], two thresholds o_{min} and o_{max} are set to calculate the soft label $\mathbf{y}_i = \frac{o_i - o_{min}}{o_{max} - o_i}$ for the i-th proposal. Note that if $\mathbf{y}_i \leq 0$, we set $\mathbf{y}_i = 0$, and we set $\mathbf{y}_i = 1$ if $\mathbf{y}_m \geq 1$. With the soft labels, we train the video moment retrieval task by two binary cross entropy losses $\mathcal{L}_{ce}(\mathbf{p}, \mathbf{y})$ and $\mathcal{L}_{ce}(\mathbf{q}, \mathbf{y})$. Taking the former as an example:

$$\mathcal{L}_{ce}(\mathbf{p}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i \log \mathbf{p}_i + (1 - \mathbf{y}_i) \log(1 - \mathbf{p}_i), \tag{8}$$

Fine-Grained Semantic Distillation. As we discussed in Section 1, the learned fine-grained information serves as complementary cues to provide an enhanced supervisory signal to the VMR model. As a result, we introduce a fine-grained semantic distillation learning, which transfers the fine-grained semantic knowledge in a form of softened matching scores. The formulation is given by:

$$\mathcal{L}_{dis}(\mathbf{p}, \mathbf{q}) = \mathcal{L}_{ce}(\sigma(\frac{\mathbf{p}}{T}), \sigma(\frac{\mathbf{q}}{T})), \tag{9}$$

where T is a temperature hyperparameter, σ is the softmax function. The softmax operation considers the distribution of proposal scores for knowledge distillation. Since $\mathbf{q}$ serves as a teacher, it is fixed and the gradient is not backpropagated through it when optimizing $\mathcal{L}_{dis}$. By distillation, the video and text encoders are required to learn from the fine-grained semantic extractor, which can be well generalized to the test stage. In our framework, the video encoder and visual feature transformation module ϕ_m is shared for both text encoder and fine-grained semantic extractor. As a result, optimizing the distillation loss (Eq. (9)) and $\mathcal{L}_{ce}(\mathbf{q}, \mathbf{y})$ can inject useful fine-grained semantics into the process of visual feature learning.

Based on the above design, the overall objective for a training video-sentence pair is formulated as:

$$\mathcal{L} = \mathcal{L}_{ce}(\mathbf{p}, \mathbf{y}) + \mathcal{L}_{ce}(\mathbf{q}, \mathbf{y}) + \lambda \mathcal{L}_{dis}(\mathbf{p}, \mathbf{q}), \tag{10}$$

where λ is a balance term.

Inference. During the test stage, we only adopt $\mathbf{p}$ for fast video moment retrieval, since calculating $\mathbf{q}$ requires additional time for fine-grained semantic extraction and feature transformation. In the learned common space, the visual moment features $\phi_m(\mathbf{m}_i)$ can be pre-calculated and stored in a gallery database, which has no impact on the test time. As a result, the computational overhead of VMR only consists of the calculation of query embeddings (text encoder and $\phi_s(\mathbf{s})$) and matching scores (Eq. (6)).

4. Experimental Results

4.1. Experimental Setup

TACos [44]. It consists of 127 videos, which contains different kitchen-related activities. The average length of the videos and moments in this dataset are 296 and 6 seconds, which makes the dataset very challenging. A standard split [14] consists of 10,146, 4,589, and 4,083 moment-sentence pairs for training, validation and testing.

ANetCap [29]. The ANetCap dataset is the largest dataset for video moment retrieval, which contains around 20,000 untrimmed action videos. It has 37,417, 17,505, and 17,031 moment-sentence pairs for training, validation, and testing, respectively. Following [65, 62], we use the first validation set for validation and the second validation set for testing.

Charades-STA [14]. The Charades-STA dataset is annotated for action recognition and localization. Charades-STA dataset contains 12,408 moment-sentence pairs in the training set and 3,720 pairs in the test set.

Evaluation Metrics. Following previous work [14, 65], we adopt the metrics R@n,IoU=m to evaluate the performance, which is defined as the percentage of at least one of the top-n predicted moments which have Intersection over Union (IoU) with ground-truth moment larger than m. Following [65, 62, 35], we set n∈$\{1, 5\}$ with m∈$\{0.1, 0.3, 0.5\}$ for TACos dataset, n∈$\{1, 5\}$ with m∈$\{0.3, 0.5, 0.7\}$ for ANetCap, and n∈$\{1, 5\}$ with m∈$\{0.5, 0.7\}$ for Charades-STA.

Implementation Details. We follow [65] to generate moment proposals, which are organized as a 2D feature map. For a fair comparison, we employ the same visual features as previous work [65]. Specifically, for the TACoS and ANetCap datasets, we adopt the C3D features [49], and for the Charades-STA dataset, we use the VGG16 features [48], C3D features [49], and I3D features [4] to evaluate the generalization on different types of features. We then apply two convolutional layers on the visual 2D feature map, where the corresponding convolutional kernel sizes are 5 and 3, respectively. We add batch norm after each convolution layer and use Tanh as the activation function. For the text encoder, we set the word embedding size as 300 and initialize it with the pretrained Glove embeddings [41]. Then a two-layer bi-directional LSTM with 512 hidden units serves for query encoding. For the fine-grained semantic extractor, the maximum numbers of verb nodes and noun nodes are set to 4 and 6 respectively. We maintain all word tokens after tokenization and truncate all text queries that have maximum 20 words. Each of the three MLPs ϕ_m, ϕ_s, and ϕ_u has one hidden layer with an output dimension of 512. ϕ_{fuse} has one convolutional and one gated convolutional layer [58] with the kernel sizes of 3 and 9, respectively. The feature dimensions D_v, D_s, D_f, and D are all set to 512. λ in Eq. (10) is set to 3. The scaling thresholds o_{max} and o_{min} are set to 0.3 and 0.7. The temperature T is empirically set to 1. We adopt a warm-up strategy, which does not optimize the distillation loss in the first 8 epochs. Our model is implemented with PyTorch 1.2.0, and we utilize Adam with a learning rate of 2×10^{-4} and a batch size of 32 for optimization. We train our model until the training loss is smooth.

Compared Methods. We compare with state-of-the-art approaches: LGI (CVPR 2020) [40], DRN (CVPR 2020) [62], 2D-TAN (AAAI 2020) [65], CBP (AAAI 2020) [51], VSLNet (ACL 2020) [64], TMLGA (WACV 2020) [45], SM-RL (CVPR 2019) [53], ACL (WACV 2019) [22], RWM-RL (AAAI 2019) [25], QSPN (AAAI 2019) [57], SAP (AAAI 2019) [9], MAN (CVPR 2019) [63], SCDM (NeurIPS 2019) [60], CTRL (ICCV 2017) [14]. Here, LGI, DRN, TMLGA, VSLNet[1], and ExCL are representative one-stage methods, while other approaches are two-stage models. In the following, the best performance is highlighted in **bold** and the second-best underline.

4.2. Comparison with State-of-the-art Methods

Overall Speed-Accuracy Analysis. The fast VMR task aims to localize target moments efficiently and accurately. During inference, the time cost of video moment retrieval is determined by two types of process: Text Encoding (*TE*) for query embedding generation and Cross-Modal Learning for moment localization (*CML*). In our proposed FVMR framework, cross-modal learning is simply implemented by vector similarity calculation in the learned common space. Table 1[2] illustrates the speed-accuracy analysis against state-of-the-art approaches, showing that our method achieves significant performance with high efficiency. Moreover, we have the following observations: (1) The time cost of *TE* is similar for all the VMR approaches (∼3 ms). As a result, *TE* is not the test-time computational bottleneck. (2) For the time cost of *CML*, our proposed FVMR method is $35\times$ to $20,000\times$ faster than state-of-the-arts, demonstrating that learning cross-modal common space is much more efficient than cross-modal interaction. (3) Overall, the proposed FVMR is a high-speed and high-quality method. Compared with the current state-of-the-art model, 2D-TAN, our proposed method is $40\times$ faster and obtains an absolute gain of 3.8% on the TACos dataset [44]. For the ANetCap and Charades-STA datasets, we also obtain superior or comparable performance in an extremely efficient fashion. (4) The one-stage approaches VSLNet and LGI also achieve favorable performance with relatively low computational costs (∼5 to 10 ms). However, they can only predict one temporal moment, which is limited in the practical retrieval scenarios. In addition, our proposed FVMR outperforms them in both speed and accuracy metrics.

Results on TACos. Table 2 summarizes the performance of different approaches on the test split of TACos. We can observe that the performance degenerates for all the methods when IoU gets higher. The proposed FVMR significantly outperforms all the other methods. Compared with the state-of-the-art method, 2D-TAN, the proposed FVMR outperforms it by an average absolute gain of 5.5%. Noticeably, the one-stage approaches such as VSLNet, DRN, and TMLGA obtain inferior performance on this dataset. The reason is that videos in TACos are often too long, which impedes the directly temporal regression of these methods.

Results on ANetCap. Table 3 reports the VMR results on the ANetCap dataset. Our proposed FVMR outperforms the state-of-the-arts such as DRN, SCDM, and 2D-TAN on most metrics. On other metrics, we achieve comparable performance. Specifically, compared with 2D-TAN, FVMR outperforms it on R@1, IoU=$\{0.3, 0.5, 0.7\}$ by gains of

[1] TMLGA and VSLNet adopts the more robust I3D features [4] for TACos and ANetCap datasets while others use C3D features [49].

[2] We evaluate all the compared methods in the same hardware environment with an NVIDIA RTX 3090.

Table 1. Speed-accuracy analysis on three datasets. **TE**: time cost of query (Text) Embedding generation. **CML**: time cost of the Cross-Modal Learning for VMR. **ALL**: The total time cost of TE and CML. We report the accuracy (**ACC**) of R@1, IoU=0.5 for comparison.

Methods	TACos				ANetCap				Charades-STA			
	TE	CML	ALL	ACC	TE	CML	ALL	ACC	TE	CML	ALL	ACC
TMLGA	**1.14**	11.37	12.51	21.65	**1.24**	8.97	10.21	33.04	**1.15**	4.37	5.52	52.02
VSLNet	3.58	5.02	8.59	24.27	3.87	4.86	8.74	43.22	3.90	4.27	8.18	54.19
LGI	-	-	-	-	1.53	7.03	8.56	41.51	1.23	4.76	5.99	**59.46**
DRN	4.67	22.13	26.81	23.17	4.86	18.46	23.32	**45.45**	4.52	12.39	16.91	53.09
CTRL	4.32	534.23	538.55	13.30	4.75	398.25	403.0	29.01	4.53	12.20	16.73	23.63
SCDM	3.65	780.0	783.65	21.17	3.27	359.76	363.03	36.75	2.97	23.77	26.07	54.44
CBP	3.17	2659.01	2662.18	24.79	2.44	522.65	525.09	35.76	2.87	266.08	268.95	36.80
2D-TAN	1.72	135.84	137.56	25.32	1.69	80.35	403.1	44.51	1.59	16.78	18.37	40.94
FVMR	3.51	**0.14**	**3.65**	**29.12**	3.14	**0.09**	**3.23**	45.00	2.86	**0.01**	**2.87**	55.01

Table 2. Comparison results on TACos.

Method	R@1			R@5		
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
TMLGA	-	24.54	21.65	-	-	-
VSLNet	-	29.61	24.27	-	-	-
DRN	-	-	23.17	-	-	33.36
CTRL	24.32	18.32	13.30	48.73	36.69	25.42
QSPN	25.31	20.15	15.23	53.21	36.72	25.30
ACL	31.64	24.17	20.01	31.64	24.17	20.01
SCDM	-	26.11	21.17	-	40.16	32.18
CBP	-	27.31	24.79	-	43.64	37.40
2D-TAN	47.59	37.29	25.32	70.31	57.81	45.04
FVMR	**53.12**	**41.48**	**29.12**	**78.12**	**64.53**	**50.00**

(1.18%, 0.49%, 0.31%) and R@5, IoU=$\{0.3, 0.5\}$ by gains of (0.58%, 0.29%).

Table 3. Comparison results on ANetCap.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
TMLGA	51.28	33.04	19.26	-	-	-
VSLNet	**63.16**	43.22	26.16	-	-	-
LGI	58.52	41.51	23.07	-	-	-
DRN	-	**45.45**	24.36	-	**77.97**	50.30
CTRL	-	14.00	-	-	-	-
QSPN	-	27.70	13.60	-	71.85	45.96
RWM-RL	-	36.90	-	-	-	-
SCDM	54.80	36.75	19.86	77.29	64.99	41.53
CBP	54.30	35.76	17.80	77.63	65.89	46.20
2D-TAN	59.45	44.51	26.54	85.53	77.13	**61.96**
FVMR	60.63	45.00	**26.85**	**86.11**	77.42	61.04

Results on Charades-STA. Since existing approaches severally adopt different types of visual features, we utilize different features for a fair comparison. As shown in Table 4, for VGG features, our method outperforms the state-of-the-art methods on most metrics. When using C3D and I3D features, our method outperforms other methods on the metrics R@1, IoU=0.5 and R@5, IoU=0.5 and also obtains comparable results on other metrics. Compared with the state-of-the-art two-stage methods, *e.g.* 2D-TAN and SCDM, our method achieves better performance on the important metric R@$\{1, 5\}$, IoU=0.5. Although VSLNet, LGI, and DRN obtain better results than ours on some metrics, they adopt a temporal location regression strategy while our method only uses fixed proposals. We leave this strategy as future work. In addition, VSLNet and LGI can only regress one temporal location for VMR, which is not suitable enough in practical scenarios.

Table 4. Comparison results on Charades-STA.

Method	Features	R@1		R@5	
		IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
SAP	VGG	27.42	13.36	66.37	38.15
SM-RL	VGG	24.36	11.17	61.25	32.08
MAN	VGG	41.24	20.54	83.21	**51.85**
2D-TAN	VGG	40.94	22.85	83.84	50.35
FVMR	VGG	**42.36**	**24.14**	**83.97**	50.15
CTRL	C3D	23.63	8.89	58.92	29.52
ACL	C3D	30.48	12.20	64.84	35.13
RWM-RL	C3D	36.70	-	-	-
QSPN	C3D	35.60	15.80	79.40	45.40
CBP	C3D	36.80	**18.87**	70.94	**50.19**
FVMR	C3D	**38.16**	18.22	**82.18**	44.96
VSLNet	I3D	54.19	35.22	-	-
LGI	I3D	**59.46**	**35.48**	-	-
DRN	I3D	53.09	31.75	89.06	**60.05**
SCDM	I3D	54.44	33.43	74.43	58.08
FVMR	I3D	55.01	33.74	**89.17**	57.24

4.3. Further Remarks

To better understand our algorithm, we conduct detailed ablation studies on the TACos dataset.

Importance of our proposed Fine-grained Semantic Extractor. In our proposed FVMR framework, the semantic role labeling module is adopted for fine-grained semantic extraction. Note that the previous approach, HGR [11], also adopted this module to perform cross-modal video retrieval. However, HGR separately utilizes each node in the semantic role graph for cross-modal matching and ignores to explicitly model the phrase-level structures, while our proposed hierarchical semantic-guided attention module can

Table 5. Ablation studies on TACos.

Model Variants	R@1			R@5		
	IoU=0.1	IoU=0.3	IoU=0.5	IoU=0.1	IoU=0.3	IoU=0.5
HGR	51.25	38.55	26.50	76.96	62.78	47.35
FVMR(w/o. SD)	48.42	38.68	27.30	74.10	60.58	48.00
FVMR($\mathbf{p}+\mathbf{q}$)	53.25	42.12	30.30	77.38	64.97	51.60
FVMR(+FeatDist)	53.67	42.56	29.34	78.52	65.01	51.12
FVMR	53.12	41.48	29.12	78.12	64.53	50.00

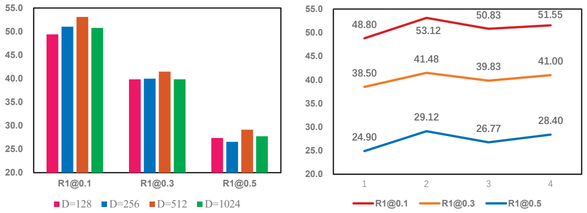
leverage the phrase-level sub-trees for more comprehensive semantic modeling. To verify the effectiveness, we adapt HGR for the VMR task, *i.e.* training the learned hierarchical textual embeddings for cross-modal query-moment matching by using Eq.(13) in [11]. HGR achieves R1@0.1,0.3,0.5 of $51.25\%, 38.55\%, 26.50\%$ on TACos. Compared with HGR, our FVMR obtains 1.16%-2.93% absolute gains.

Importance of the Fine-grained Semantic Distillation. To investigate the effect of the semantic distillation, we design a baseline FVMR(w/o. SD) that removes the sematic distillation loss (Eq. (9)) during model training. From Table 5 we can find that our full model FVMR outperforms the FVMR(w/o. SD) by absolute gains of (4.70%, 2.80%, 1.82%) on R@1 metric and (4.02%, 3,95%, 2.0%) on R@5 metric. The results validate the importance of transferring knowledge from fine-grained semantic structures. Another observation is that FVMR(w/o. SD) obtains better performance than 2D-TAN. This is because the semantic information can still facilitate the learning of robust moment features by optimizing the video moment retrieval loss $\mathcal{L}_{ce}(\mathbf{q}, \mathbf{y})$.

How about Fusing $\mathbf{p}$ and $\mathbf{q}$? In our framework, two types of matching scores can be obtained: (1) the matching score $\mathbf{q}$ from the fine-grained semantic extractor, which serves as the teacher in distillation. (2) the matching score $\mathbf{p}$, which is utilized for fast moment retrieval. To investigate the performance of fusing $\mathbf{p}$ and $\mathbf{q}$ in VMR, we design a baseline FVMR($\mathbf{p}+\mathbf{q}$), which employs the average of $\mathbf{p}$ and $\mathbf{q}$ as the final results. From Table 5 we can observe that FVMR($\mathbf{p}$+$\mathbf{q}$) achieves higher performance because it leverages both types of information from common space learning and cross-modal interaction. However, calculating $\mathbf{q}$ costs about 35 ms in our experiments, which is not suitable for fast VMR. As a result, it is advisable to directly utilize $\mathbf{p}$ during inference, which also has favorable performance.

Effect of D. We investigate the effect of D, which is the dimension of the learned moment-query common space. Intuitively, a larger D will result in a more discriminative and comprehensive common space. However, a too-large D leads to a high computational burden and may result in model overfitting. As shown in Figure 3(a), a moderate value of D obtains the best performance.

Number of Bi-LSTM Layers. Theoretically, a deeper network will enhance the feature learning process, thus improving the model robustness. However, as shown in Figure 3(b), we observe that adding more layers does not boost the performance. The reason is that stacking many layers can result in the overfitting problem. In addition, more layers lead to higher computational costs. As a result, we set the number of layers to a moderate value of 2.



(a) Common space dimension D (b) Number of Bi-LSTM layers

Figure 3. Parameter analysis of common space dimension D and the number of Bi-LSTM layers.

Regularization on the Textural Feature Space. There is a simple alternative that replaces the matching score distillation with textual feature discrepancy minimization, i.e., performing feature distillation between $\mathbf{u}$ and $\mathbf{s}$. In our experiments, we find that using the feature distillation achieves similar performance. Inspired by this, we design a baseline FVMR(+FeatDist), which jointly uses both types of distillation in our framework. As shown in Table 5, the baseline obtains a performance improvement (0.2%-1.1%). Intuitively, the two constraints can be viewed as distribution-level distillation and single feature-level distillation, which complement each other. We leave the exploration of advanced distillation strategies as our future work.

5. Conclusions

This paper introduces the fast video moment retrieval task that aims to retrieve moments efficiently and accurately. To this end, a fine-grained semantic distillation framework is designed, which learns a moment-query common space by transferring knowledge from additional semantic structures. During inference, efficient vector operations are conducted in the common space for fast video moment retrieval. In the future, to improve the test speed and save the storage space of video features, cross-modal hashing and model compression strategies could be considered. In addition, for the sake of higher performance, we aim to leverage a more effective fusion module ϕ_{fuse} to learn semantic feature interaction, such as co-attention and transformer. We believe that fast VMR will significantly facilitate the development of relevant applications in our daily life.

Acknowledgements

This work was supported by the National Key Research & Development Plan of China under Grant 2020AAA0106200, in part by the National Natural Science Foundation of China under Grants 62036012, 61721004, 62072286, 61720106006, 61832002, 62072455, 62002355, U1836220, and U1705262, in part by the Key Research Program of Frontier Sciences of CAS under Grant QYZDJSS-WJSC039, and in part by Beijing Natural Science Foundation (L201001).

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*, 2017. 1, 2, 3

[2] Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *ACM MM*, 2020. 2, 3

[3] Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. Adversarial video moment retrieval by jointly modeling ranking and localization. In *ACM MM*, 2020. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 3, 6

[5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018. 2

[6] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *AAAI*, 2019. 2, 3

[7] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 1

[8] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *ECCV*, 2020. 2

[9] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *AAAI*, 2019. 2, 6

[10] Shaoxiang Chen and Yu-Gang Jiang. Hierarchical visual-textual graph for temporal activity localization via language. In *ECCV*, 2020. 3

[11] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020. 3, 4, 7, 8

[12] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS Workshops*, 2014. 1, 4

[13] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 2, 3, 5, 6

[15] Junyu Gao and Changsheng Xu. Learning video moment retrieval without a single annotated video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[16] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. A unified personalized video recommendation via dynamic recurrent neural networks. In *MM*, pages 127–135. ACM, 2017. 2

[17] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling. In *MM*, pages 690–699, 2018. 2

[18] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *CVPR*, pages 4649–4659, 2019. 2

[19] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 2

[20] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[21] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu. Deep relative tracking. *IEEE Transactions on Image Processing*, 26(4):1845–1858, 2017. 2

[22] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *WACV*, 2019. 3, 6

[23] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *ACL*, 2019. 2

[24] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. Tripping through time: Efficient localization of activities in videos. In *BMVC*, 2020. 2, 3

[25] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019. 3, 6

[26] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *EMNLP*, 2018. 2, 3

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 3

[28] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *ICMR*, 2019. 3

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

[30] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 1

[31] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 3

[32] Zhijie Lin, Zhou Zhao, Zhu Zhang, Zijian Zhang, and Deng Cai. Moment retrieval via cross-modal interaction networks with query reconstruction. *IEEE Transactions on Image Processing*, 29:3750–3762, 2020. 3

[33] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, 2018. 3

[34] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM MM*, 2020. 2, 3

[35] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, 2018. 1, 2, 6

[36] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM MM*, 2018. 1

[37] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP-IJCNLP*, pages 5147–5156, 2019. 2

[38] Esa Rahtu Mayu Otani, Yuta Nakahima and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*, 2020. 1

[39] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 3

[40] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 1, 2, 3, 6

[41] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6

[42] Sisi Qu, Mattia Soldan, Mengmeng Xu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. *arXiv preprint arXiv:2011.10132*, 2020. 1, 2

[43] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *ACM MM*, 2020. 2, 3

[44] 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. 2, 5, 6

[45] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020. 2, 6

[46] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 2018. 1

[47] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019. 4

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[49] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 3, 6

[50] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *ACM MM*, 2020. 2

[51] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, 2020. 2, 3, 6

[52] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. Learning coarse-to-fine graph neural networks for video-text retrieval. *IEEE Transactions on Multimedia*, 2020. 2

[53] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019. 3, 6

[54] Aming Wu and Yahong Han. Multi-modal circulant fusion for video-to-language and backward. In *IJCAI*, 2018. 2

[55] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, 2020. 2

[56] Ziyue Wu, Junyu Gao, Shucheng Huang, and Changsheng Xu. Diving into the relations: Leveraging semantic and visual structures for video moment retrieval. In *ICME*, 2021. 3, 4

[57] 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*, 2019. 1, 3, 6

[58] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 6

[59] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 2

[60] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, 2019. 2, 3, 6

[61] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 2

[62] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 1, 2, 5, 6

[63] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 1, 2, 3, 6

[64] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 1, 6

[65] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 1, 2, 3, 5, 6

[66] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *ACM MM*, 2019. 3

[67] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, 2019. 1, 3

[68] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *CVPR*, 2020. 1