Frame-wise Cross-modal Match for Video Moment Retrieval

Haoyu Tang, Jihua Zhu, Meng Liu, *Member*, *IEEE*, Zan Gao, and Zhiyong Cheng

Video moment retrieval targets at retrieving a golden moment in a video for a given natural language query. The main challenges of this task include 1) the requirement of accurately localizing (i.e., the start time and the end time of) the relevant moment in an untrimmed video stream, and 2) bridging the semantic gap between textual query and video contents. To tackle those problems, early approaches adopt the sliding window or uniform sampling to collect video clips first and then match each clip with the query to identify relevant clips. Obviously, these strategies are time-consuming and often lead to unsatisfied accuracy in localization due to the unpredictable length of the golden moment. To avoid the limitations, researchers recently attempt to directly predict the relevant moment boundaries without the requirement to generate video clips first. One mainstream approach is to generate a multimodal feature vector for the target query and video frames (e.g., concatenation) and then use a regression approach upon the multimodal feature vector for boundary detection. Although some progress has been achieved by this approach, we argue that those methods have not well captured the cross-modal interactions between the query and video frames.

In this paper, we propose an Attentive Cross-modal Relevance Matching (ACRM) model which predicts the temporal bounders based on an interaction modeling between two modalities. In addition, an attention module is introduced to automatically assign higher weights to query words with richer semantic cues, which are considered to be more important for finding relevant video contents. Another contribution is that we propose an additional predictor to utilize the internal frames in the model training to improve the localization accuracy. Extensive experiments on two public datasets TACoS and Charades-STA demonstrate the superiority of our method over several state-of-the-art methods. Ablation studies have been also conducted to examine different modules in our ACRM model.

***Index Terms*—Video Moment Retrieval, Cross-modal Retrieval, Moment Localization, Frame-wise Matching.**

I. Introduction

VISUAL-language understanding plays an important role in developing artificial intelligence in human-computer interactions [1], [2], [3], [4], [5], [6], [7]. Particularly, video retrieval has drawn significant attention over the past decades. Given a text query, such as "person put a notebook in a bag", the goal of video retrieval is to find videos that contain relevant content with respect to the query. To watch the specific video clip which is relevant to the query, we need to browse the video to localize the relevant part in a video, which could take hours, especially in the surveillance video scenarios. Therefore, it is important to find relevant video clips with accurate temporal boundaries (i.e., the start time and the end time) for a given query, which is the so-called video moment retrieval task. This is a recently emerged research topic and has attracted increasing attention due to its practical value [2], [3].

Particularly, the target of moment retrieval is to precisely localize a moment in the untrimmed video whose content is in accordance with the given arbitrary natural descriptions [1], as illustrated in Figure 1. Based on the experience of approaches in video retrieval, early approaches follow a two-step manner, i.e., generating the moment candidates via the temporal sliding window strategy and then matching them with the query in a common cross-modal space [1], [8], [9], [10], [11]. Because the desired moments can be of varying lengths, various sizes of sliding windows need to be employed to generate numerous overlapping segments to match with the query. Therefore, this type of method is cumbersome and resource-consuming.

To reduce the number of moment candidates, several methods have been developed, such as uniformly sampling segments in the video [12], while others leveraging the segment proposal network [13] to automatically generate moment candidates that most likely contain the potential activities [14], [15]. However, those methods still need to first generate moment candidates and then match them with the query, resulting in inferior efficiency. Moreover, although these approaches refine the boundaries of the selected moment candidates, the performance will still be far from satisfactory when the selected candidate has a little overlap with the ground-truth moment.

To overcome these drawbacks, several one-step retrieval methods without the requirement of generating video segments have been proposed [16], [17]. Specifically, the cross-modal feature of the query and each video frame is used to predict the probability of the frame to be the starting or ending frame of the target moment. The concatenation of the query feature and frame feature is often used in the previous method to generate the cross-modal feature [18], [19], [20], [21]. Although significant improvement has been achieved, there are still several limitations. First, in those methods, the video features and query features are extracted via separate networks. As a result, the video feature and query feature are from different feature spaces, but they are directly concatenated for the next step prediction. In addition, the concatenation of these features

Z. Cheng is the corresponding author.

H. Tang is with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China, and with the School of Software Engineering, Xian Jiaotong University, Xian 710049, China. (e-mail: tanghao258@stu.xjtu.edu.cn)

J. Zhu is with the School of Software Engineering, Xian Jiaotong University, Xian 710049, China. (e-mail: zhujh@stu.xjtu.edu.cn)

M. Liu is with Shandong Jianzhu University, Jinan 250101, China. (e-mail: mengliu.sdu@gmail.com)

Z. Gao, Z. Cheng are with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China. (e-mail: tanghao258@stu.xjtu.edu.cn, zangaonsh4522@gmail.com, jason.zy.cheng@gmail.com)

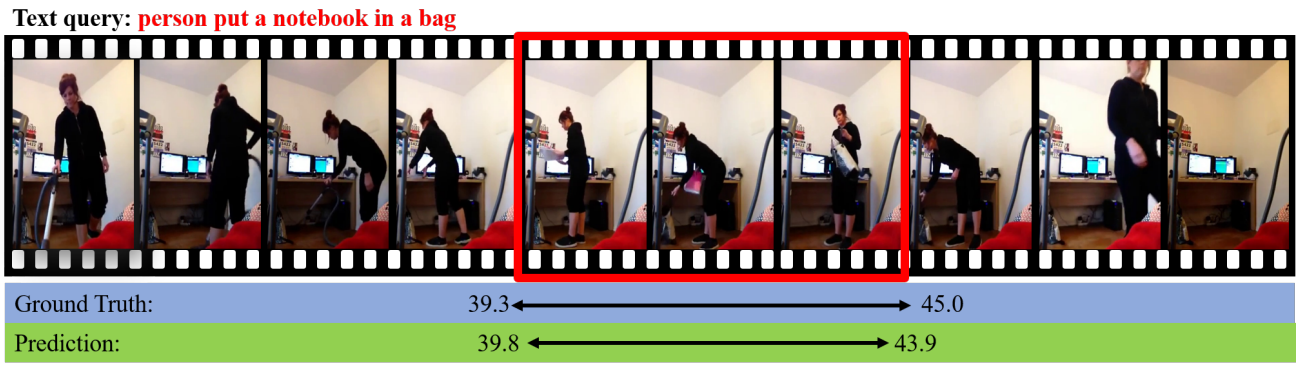


Fig. 1. Moment retrieval aims to localize the temporal boundaries with a start time (39.3s) and an end time (45.0s) of a desired moment in the red box, corresponding to the given query "Person put a notebook in a bag.".

cannot well capture the interactions between the query and video content. Besides concatenation, other methods use the query feature to attend the video frame to generate a weighted video feature for boundary prediction [21], [22]. Similarly, they fail to model the fine-grained interactions between video frames and query words, which are important to guide the localization. In addition, the problem of different feature spaces has not been tackled either in those methods. Intuitively, the relevant video frames should be closer to the query than the irrelevant ones in the shared feature space, which is a common assumption in the cross-modal retrieval problem. From this perspective, the concatenation or the attended video features in the above methods cannot well model the interaction between the frame feature and query feature.

Moreover, most existing methods mainly focus on predicting the temporal boundaries and overlook the internal frames (i.e., the ones between the start and end frames), which also contain valuable information [20]. For the query person put a notebook in a bag as illustrated in Figure.1, the desired moment in the video is the process of the person picking up the book and putting it into the bag. Notably, as the camera is locked in a specific view, the notebook has not even appeared outside the desired moment and thus the semantic information in all frames of the desired moment is similar. Under such circumstances, the cross-modal features extracted by the model on these frames within the moment become very similar, resulting in indistinguishable prediction scores between the internal frames and the boundary frames, which is not beneficial for the judgment of the boundaries. If the model also predicts whether a frame belongs to the desired moment, the information of internal frames can be leveraged to enhance the prediction of boundaries.

Based on the above considerations, we propose a novel moment retrieval model called Attentive Cross-modal Relevance Matching (ACRM), which directly predicts the relevant moment boundaries without the requirement of pre-processed candidates of video clips. Specifically, we process each frame of a target video and embed the frame feature into the same space with the query feature. Rather than employing the concatenation or the attended video features, we use an interaction function to model the interactions between the frame and the query for boundary prediction. In addition, a multimodal attention module is introduced to estimate the importance of each word in the query. This can not only enhance the cross-modal match between video frames and queries but also exploit the fine-grained frame-query interactions. Besides, we incorporate the prediction of internal frames as element moment frames into the objective function, which can effectively improve the boundary prediction accuracy. Extensive experiments have been conducted on two benchmark datasets Charades-STA [1] and TACoS [23]. The results show that our model can consistently outperform all the competitors by a large margin.

In summary, the main contributions of this work are threefold:

- We highlight the importance of modeling the interactions between the cross-modal features for video moment retrieval and propose a novel ACRM model. In particular, an attention mechanism and a similarity function are integrated into ACRM to model the interactions between the video frame and query features.
- To fully leverage the information containing in the internal frames of the moment, we add an extra predictor to estimate the probability of a frame to be the internal frame, which is proven to be effective in our experiments.
- We conducted extensive experiments to evaluate the performance of our proposed model by comparing it to several state-of-the-art methods. We also analyze the effectiveness of each module in our model by ablation studies. Our code has been released for the reproduction of the experiments.[1]

II. Related Work

A. Moment Retrieval

Finding the desired moment in an untrimmed video according to a sentence query is a challenging task due to the requirement of cross-modal understanding. To accomplish this task, early methods follow a two-step manner, which generates moment candidates first and then matches them with the query [2] to find the relevant ones. For example, Gao et al. [1] presented a Cross-modal Temporal Regression Localizer (CTRL), which generates moment candidates via a temporal

[1] https://github.com/tanghaoyu258/ACRM-for-moment-retrieval

sliding window method, and then encodes those candidates and the sentence query into the same space to find the relevant candidates by matching the candidates with the query in the space. Following this framework, Liu et al. proposed a ROLE model [8] which uses a language temporal attention module to learn sentence representation, and an ACRN model [9] which adopts a memory attention network to capture the contextual information of the moments. Although the above methods achieve better performance with advanced representation learning methods, they are still resource-consuming due to the sliding window strategy.

To overcome this limitation, some research efforts have been dedicated to reducing the number of generated temporal moment candidates [12], [24]. For instance, inspired by the R-C3D [13] model which was designed for action localization in the video, Xu et al. [14] employed a segment proposal network [13] to generate the varied-length moment candidates. Later on, they [15] presented a multilevel language and vision integration model which incorporated sentence features to generate the attended moment candidates. Wang et al. [25] proposed a temporal grounding model that explores the interactions between video sequence and sentence to simultaneously score multiple candidates at each time step. Zhang et al. [26] employed a downsampling strategy to reduce the number of candidates obtained from a two-dimensional temporal map.

Recently, researchers attempt to directly localize the desired moments without the requirement of generating candidate first. Several methods have been proposed in this direction. For instance, the reinforcement learning (RL) strategy has been adopted in moment retrieval [16], [17]. In general, the RL-based methods progressively update the temporal boundaries over the entire video to locate the desired moment for a given query. Meanwhile, another research line is to predict the probabilities of each frame to be the boundary frame based on the cross-modal feature of the query and video frame. For example, Yuan et al. [21] proposed to concatenate the attended query feature and the video frame features and then regress the temporal interval to the boundaries for each frame. The ExCL model [18] used a regression method upon the concatenation of the query and video frame feature to predict the boundary frame. Following this work, Rodriguez et al. [22] predicted the boundary based on the attentive video features, which are attended by the query feature. Chen et al. [19] and Zhang et al. [20] concatenated the query feature, video feature, and their similarity together for the subsequent prediction of the temporal boundaries. We argue that the above method (i.e., concatenation of query and video frame feature, and the attended video features) cannot well capture interactions between cross-modal features. In this work, we encode the query and video frame into the same feature space to obtain a similarity vector, upon which a prediction model is used to detect the boundary. In addition, we adopt an attention mechanism to identify meaningful query words and add an internal frame predictor into the objective function to enhance the boundary detection performance. It is worth noting that some methods also exploit the information of the internal frames. Rodriguez et al. [22] proposed to highlight the temporal attention weights across the internal frames, and Zhang et al. [20] temporally extended the region of the internal frames to include more video context. Different from those methods by extending the region of the internal frames, our model exploits the information of internal frames by using an additional predictor in the objective function.

B. Temporal Action Localization

Temporal action localization is a similar task that also needs to localize the boundary of required moments in videos. The difference is that they have a pre-defined action list and the required moments are action instances in the list. As this task has pre-defined concepts, most methods are supervised. A general pipeline is to first generate candidate clips containing activities and then use the pre-trained action classifiers to detect the action. For instance, Gao et al. [27] jointly classified the action proposals and fine-tuned their temporal boundaries in a temporal unit regression network, and Xu et al. [13] introduced a Region Convolutional 3D model (R-C3D) to generate temporal candidates containing activities, which were finally classified. Ma et al. [28] proposed to incorporate the predicted scores from a temporal LSTM over the detection span. There are also approaches designed in a weakly-supervised fashion. Wang et al. [29] integrated the classification module and the selection module to predict the action proposals and then select the most probable ones, respectively. Paul et al. [30] leveraged an attention-based module along with multiple instances learning to learn pairwise video similarity constraints for localization and classification. Despite great progress has been achieved for the task of temporal action localization, those methods are not suitable for the moment retrieval task, because we do not have knowledge about the queries. In fact, video moment retrieval is a more complex task, as the query can be of arbitrary lengths and a variety of concepts, which cannot be predefined at all.

C. Sentence-based Video Retrieval

Another closely related task is sentence-based video retrieval, which aims to search the most relevant video from a video set. Many methods regard it as a ranking problem by mapping the videos and sentences into a common space [31], [32], [33], [34], [35]. A general approach is to feed the video into a pre-trained CNN model to extract frame features, which are aggregated into a video feature by mean-pooling. Such processing leads to inefficiency in learning a common embedding space since the temporal cues cannot be captured. To tackle this problem, Dong et al. [36] proposed to encode global and local pattern information for both video and text ends. Lin et al. [34] presented to retrieve the video by matching a semantic graph from parsed sentence descriptions and the visual concepts in the video. Besides, Mithun et al. [37] proposed to learn two different embedding spaces to obtain temporal and appearance information. The advancement of sentence-based video retrieval is beneficial to the development of video moment retrieval task, especially the techniques of matching query and video contents, as both tasks need to find relevant video content for the targeted query. The difference is that video moment retrieval needs to precisely localize the boundary of the relevant video clips, which is much more

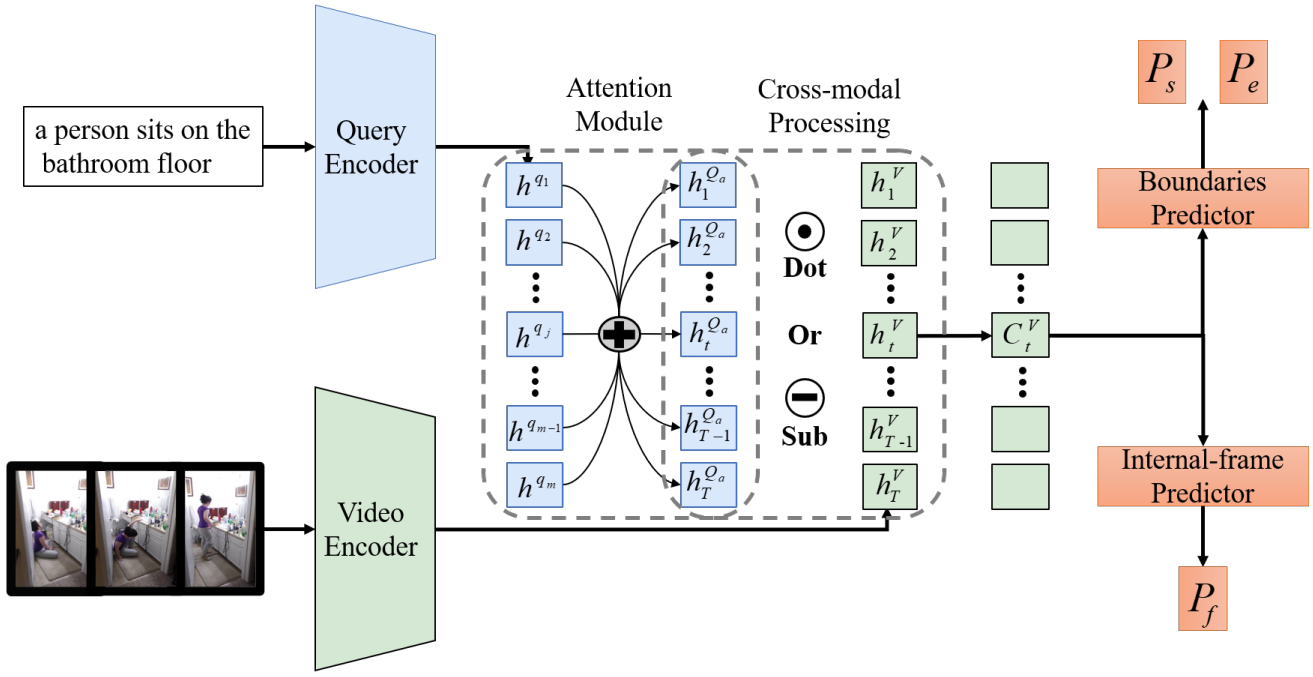


Fig. 2. An illustration of the detailed ACRM model. It comprises of four components: two encoders to extract the video frame features and textual embeddings separately, an attention module to generate frame-specific query representation, a processing to calculate the cross-modal interactions, and a two-branch predictor to estimate the temporal boundaries.

challenging than just find the relevant videos from a video set.

III. The Proposed Model

A. Preliminaries

Before formally describing our model in detail, we first introduce the primary notations in this paper. Given a video $V = \{v_t\}_{t=1}^{T}$, where v_t represents the image frame at time t and T is its length, and a sentence query $Q = \{q_j\}_{j=1}^{m}$ with m words, our goal is to find the relevant moment in the video by identifying the accurate start and end frames. Formally, this problem can be formulated as a mapping function:

$$L_\theta : (V, Q) \rightarrow (t^s, t^e), t^s < t^e \tag{1}$$

where t^s and t^e represent the start time and the end time of a golden video moment, respectively. The proposed model is trained in an end-to-end fashion on the training set, which contains K instances. Each instance is a video-query-boundaries tuple $\{V, Q, \tau^s, \tau^e\}$, where the query Q is associated with the ground-truth start and end time point τ^s and τ^e in the video V. During the evaluation stage, given an unseen video-query pair $\{V_e, Q_e\}$, the goal is to predict the temporal boundaries $\{t^{s'}, t^{e'}\}$ in the video V_e.

B. Our Model

Figure 2 shows an overview of our model, which consists of four components: 1) **Feature extraction** module extracts the video frame and query features by two separate networks; 2) **Cross-modal interaction** module models the interaction between the features of two modalities; 3) **Attention Module** attentively fuses the cross-modal features to generate the frame-specific query representation; and 4) **Prediction module** estimates the probability of a frame to be a boundary frame and an internal frame. In the next, we introduce the four modules in sequence.

1) Feature Extraction

In our model, the features of video frames and queries are extracted by two separate networks.

Video feature extraction. We apply the off-the-shelf visual feature extractors to extract the features of each frame for an untrimmed video $V = \{v_1, \ldots, v_T\}$, such as the I3D [38] or the C3D extractor[39]. The BiLSTM [40] is employed here to sequentially process the extracted visual feature because it encodes video sequence from bi-directions. In this way, every frame representation will be affected by its contiguous frames spontaneously to augment the incorporated contextual information. Specifically, every hidden state of the forward and the backward LSTM are concatenated together to obtain the new representations. The video encoder is defined as follows:

$$\begin{aligned} \mathbf{f}_t &= F(v_t) \\ \mathbf{h}_t^V &= BiLSTM(\mathbf{f}_t, \mathbf{h}_{t-1}^V) \end{aligned} \tag{2}$$

where $F(\cdot)$ represents the visual extractor, and $\mathbf{f}_t$ is the extracted feature of the t-th frame. $\mathbf{h}^V \in R^{T \times d}$ is the hidden state of the video and d is the dimension of the feature vector.

Query feature extraction. As for the sentence query Q, the pre-trained glove [41] embedding is used to transform the query words $Q = \{q_1, \ldots, q_m\}$ into embeddings $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_m\}$, i.e. $\mathbf{S} = glove(Q)$. Note that here we can use any other word embedding approaches such as Skip-Thought [42]. With the sequential word embeddings S as input, a BiLSTM network is used to encode the sentence into the representation $\mathbf{h}^Q = \{\mathbf{h}^{q_1}, \mathbf{h}^{q_2}, \cdots, \mathbf{h}^{q_m}\} \in R^{m \times d}$,

where d is the dimension of the hidden state, which is the same as the feature vector of the video frame. Compared to glove methods, the BiLSTM could comprehensively encode the context information of the whole sentence.

2) Cross-modal Interaction

Previous methods like [18], [8] directly concatenate the extracted video frame feature and query feature for the next step prediction. However, the frame feature and query feature are extracted by different networks. On one hand, the learned features based on two networks are from different spaces, the direct concatenation is problematic. On the other hand, the values of frame feature and that of query features might be in different range. As a result, the concatenated feature will be of great variance, which increases the difficulty of learning a good prediction model in the next step. Besides, the concatenation cannot well capture the interaction between the two modalities. In cross-modal retrieval, a successful approach for the cross-modal match is to compute similarity (such as element-wise multiplication or Euclidean distance) between the feature vectors of two modalities in the same feature space. Motivated by this observation, in this work, we propose to use different interaction functions to model the interaction between the frame and query features.

Specifically, we first embed the video frame feature and the query feature into a common space, in which the interactions between the frame and query are modeled. Formally, the video and query features are embedding to the same space by a transformation matrix:

$$\begin{cases} \hat{\mathbf{h}}^{Q} = \mathcal{N}\left(\mathbf{W_q}\mathbf{h}^{Q_a} + \mathbf{b_q}\right) \\ \hat{\mathbf{h}}^{V} = \mathcal{N}\left(\mathbf{W_v}\mathbf{h}^{V} + \mathbf{b_v}\right) \end{cases} \tag{3}$$

where $\mathbf{W_q}$ and $\mathbf{W_v}$ are trainable weight matrices, and $\mathbf{b_q}$ and $\mathbf{b_v}$ are bias vectors. $\mathbf{h}^{Q_a} \in R^{T\times d}$ is the query feature calculated through the attention module, which we will explain in the section III-B3. $\mathcal{N}(\cdot)$ is an operation to transform the values of the query and video frame features into the same range (e.g., [-1, 1]). Different strategies can be used here to achieve the goal. Here, we test two popular approaches. One is to use an activation function which are widely used in neural networks and the **tanh** activation function is used here; the other one is the normalization method and the Gauss distribution normalization is adopted. Finally, the interaction between the frame and query are modeled:

$$\mathbf{C^V} = f(\hat{\mathbf{h}}^{V}, \hat{\mathbf{h}}^{Q_a}) \tag{4}$$

where $f(\cdot)$ is the interaction function. Two simple yet effective interaction functions are explored in this work: **element-wise multiplication** and **subtraction**.

3) Attention Module

Since the localization of moment boundaries needs the frame feature and query feature, the integration of obtained word embeddings into a discriminative representation of the query is crucial. A widely used strategy is the mean pooling of the embedding from all the hidden states. A limitation of this method is that it treats each word in the query equally. However, it is common that some words convey more information to localize the relevant frames. For example, for the query "the black car is arriving", the word "arriving" conveys the crucial temporal information, which is thus more helpful on identifying the desired moments. The mean pooling cannot distinguish the different importance of words on identifying the relevant moments in the video.

To tackle this issue, we design an attention module to generate frame-specific query representation to fully explore the relations between the query and the video context. Specifically, it employs the t-th frame feature $\mathbf{h}_t^V$ to adaptively attend all word features in the query $\mathbf{h}^Q \in R^{m\times d}$ to obtain a summarized query feature $\mathbf{h}_t^{Q_a}$. Accordingly, the attention weight represents the relevance between the t-th frame and each words in the query. The attention module is detailed as:

$$r_{tj} = \mathbf{w}_r^T \cdot \tanh\left(\mathbf{W}_s\mathbf{h}^{q_j} + \mathbf{W}_v\mathbf{h}_t^V + \mathbf{b}_r\right) \tag{5}$$

$$\beta_{tj} = \frac{\exp\left(r_{tj}\right)}{\sum_{k=1}^{m}\exp\left(r_{tk}\right)} \tag{6}$$

where the weight matrix $\mathbf{W}_s$ and $\mathbf{W}_v$ encode the hidden state of the t-th frame and the hidden state of the j-th word into a common space to compute r_{tj}, which is fed in Eq.6 to obtain the normalized attention weight β_{tj}. $\mathbf{w}_r^T$ is a trainable vector. The t-th query feature is summarized as:

$$\mathbf{h}_t^{Q_a} = \sum\nolimits_{j=1}^{m}\left(\beta_{tj}\cdot\mathbf{h}^{q_j}\right) \tag{7}$$

After each frame is processed, the obtained frame-specific query features are concatenated as $\mathbf{h}^{Q_a} \in R^{T\times d}$, which is the frame-specific query feature in Eq. 3.

4) Prediction

To estimate the start and end frame of a specific moment, the prediction module processes the obtained cross-modal features and outputs the prediction vector with the same length T as the input frames. In previous methods, the predictors only consider the prediction of the start frame and end frame by maximizing the scores of the ground-truth boundaries. They have not exploited the information of internal frames between the boundaries in the model training. In fact, those internal frames contain rich information to help the localization of the temporal boundary. For example, if an internal frame matches the query well, it should be in the desired moment video clips.

Based on the above considerations, we integrate an additional internal frame predictor to estimate the probability of a frame to be the internal frame. The prediction of a frame to be a start frame, end frame, or the internal frame of the desired moment follows the same pipeline with separate prediction networks. Many prediction functions can be used here, such as the Multi-layer Perceptron (MLP) with a regression loss or an LSTM with a classification loss. The tied LSTM predictor with the classification loss is selected as the backbone of our prediction module because of its simplicity. In the next, we take the prediction of the start frame as an example to describe the process. Specifically, for the frame sequence, the cross-modal features (obtained by Eq. 4) are first processed by an LSTM, whose outputs are then fed into an MLP network. Finally, the softmax function is used to predict its probability to be the start frame. Formally, given the cross feature of the frame-query pair $\mathbf{C^V}$,

$$\mathbf{h}_t^P = BiLSTM(\mathbf{C}_t^V, \mathbf{h}_{t-1}^P) \tag{8}$$

where $\mathbf{h}_t^P$ is the t-th hidden state of the BiLSTM. All the obtained hidden states are concatenated as $\mathbf{h}^P \in R^{T\times d}$, which is fed in the MLP layer as follows:

$$\mathbf{e}_s = MLP\left(\mathbf{h}^P\right) \tag{9}$$

$$\mathbf{P}_s = softmax(\mathbf{e}_s) \tag{10}$$

where $\mathbf{e}_s$ is the output vector of MLP and $\mathbf{P}_s \in R^T$ is the start frame probabilities vector.

C. Learning

The objective function consists of two parts. Firstly, a classification loss, which encourages the correct start and end frame to have a larger prediction score by maximizing the log-likelihood of the ground-truth boundaries, is employed for the boundary detection. Secondly, for each video-query pair, we would like the frames between the boundaries have a larger prediction score than the outside ones. This is achieved by maximizing the averaged log-likelihood of the internal frames.

$$L_c = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\mathbf{P}_s\left(\tau_i^s\right)\right) + \log\left(\mathbf{P}_e\left(\tau_i^e\right)\right) \tag{11}$$

$$L_I = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=s}^{e}\log\left(\mathbf{P}_f\left(\tau_i^j\right)\right)/\left(\tau_i^e - \tau_i^s\right) \tag{12}$$

where $\mathbf{P}_s$, $\mathbf{P}_e$, $\mathbf{P}_f$ is the start frame, end frame, and the internal-frame probabilities, respectively. τ_i^s and τ_i^e represents the ground-truth temporal boundaries of the i-th video-query pair. The overall loss function is summarized as:

$$L = L_c + \lambda L_I \tag{13}$$

where λ balances the two losses. Note that the internal-frame predictor is used to improve the learning process in our model, and it is only used in the training stage.

In the test stage, the boundaries are determined only by the boundary prediction as follows:

$$\begin{aligned} t^s, t^e &= \arg\max_{t^s, t^e} \mathbf{P}_s\left(t^s\right)\mathbf{P}_e\left(t^e\right) \\ &= \arg\max_{t^s, t^e} \mathbf{e}_s\left(t^s\right) + \mathbf{e}_e\left(t^e\right) \\ s.t.\, t^s &\le t^e \end{aligned} \tag{14}$$

IV. Experiment

Comprehensive experiments have been conducted on two datasets Charades-STA and TACoS. The settings for all other approaches are the same as what they have reported in their papers, such as the dataset splits, and the hyper-parameters.

A. Experiment Setup

1) Dataset

Charades-STA [1]: This dataset was extended from the original Charades dataset in [43], which is mainly used for temporal activity localization task and only contains the video-level description. To accommodate the moment retrieval task, Gao et al. [1] generated the sentence-clip annotations by decomposing the provided descriptions into shorter parts, which were assigned to the clips and further manually verified by annotators. We follow the experimental setting defined in [1], in which the training set has 12408 video-sentence pairs and the testing set has 3720 video-sentence pairs. The videos are 30 seconds long on average.

TACoS [23]: This dataset is collected from the MPII Cooking Composite Activities dataset [23], which contains 127 long videos in the cooking scenarios. Following the dataset split as in Gao et al [1], 10,146, 4,589, and 4,083 clip-sentence pairs are employed for training, validation, and testing, respectively. In the dataset, the timespans of the labeled moments are short and the overlapping between the moments is often large, which makes it a challenging dataset.

2) Implementation Details

Each sentence is tokenized by Stanford CoreNLP [44] and then the pre-trained 300-dimensions glove embeddings [41] are adopted to obtain the word-level representations. The vocabulary size is set as 3720 and 1438 for Charades-STA and TACoS, respectively. For the video frame representations, the I3D features [38] were extracted for both the Charades-STA dataset and the TACoS dataset. Note that the parameters of the I3D extractor and the glove embeddings are fixed during the training. The size of the visual and query bidirectional LSTM encoders are 256 dimensions, and the MLP predictor consists of a single 256-dimension hidden layer with **tanh** as the activation function. We train our model with a batch size of 64 and adopt the early-stopping strategy for both two datasets. The Adam optimizer is adopted with learning rate of 0.001. Besides, λ is empirically set as 0.7, 1.1 for the Charades-STA and TACoS dataset, respectively. Dropout is adopted in LSTMs to prevent over-fitting and the dropout ratio is set to 0.5.

3) Evaluation Metrics

The standard evaluation metrics "R@n, IoU=m" [45] and "mIoU" [10] are used for evaluation. To be specific, the Intersection over Union (IoU) between the predicted and ground-truth temporal boundaries of the top-n retrieval results for each query is computed, and "R@n, IoU=m" is the percentage of instances that have at least one in top-n retrieval results with IoU larger than the threshold m. "mIoU" represents the mean IoU of top-**1** result across all queries.

B. Comparison with State-of-the-Arts

1) Baselines

The proposed model is compared with several state-of-the-art methods. For a fair comparison, we directly copy the reported results from the original papers of those methods. The considered competitors are listed as follows:

- CTRL [1]: This method considers the mean-pooled video contexts of the moment candidate generated by a sliding window, and matches the obtained feature with the query.
- SM-RL [46]: This RL-based model adaptively observes the video sequence and then matches the video content with the query.
- ABLR [21]: This model incorporates a cross-modal co-attention mechanism to learn video and query attentions, which are used to localize the moment.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND THE STATE-OF-THE-ARTS ON CHARADES-STA DATASET. THE SYMBOL '*' MEANS THE I3D FEATURES ARE ADOPTED.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	mIoU
CTRL(ICCV'2017)	-	23.63	8.89	-
SM-RL(CVPR'2019)	-	24.36	11.17	-
ABLR(AAAI'2019)	-	24.36	9.01	-
SAP(AAAI'2019)	-	27.42	13.36	-
ACL(WACV'2019)	-	30.48	12.2	33.84
MLVI(AAAI'2019)	54.70	35.60	15.80	-
TripNet(CVPR'2019)	51.33	36.61	14.50	-
CBP(AAAI'2020)	-	36.80	18.87	35.74
GDP(AAAI'2020)	54.54	39.47	18.49	-
2D-TAN(AAAI'2020)	-	39.81	23.31	-
SCDM*(NIPS'2019)	-	54.44	33.43	-
ExCL*(EMNLP'2019)	65.10	44.10	22.60	-
DRN*(CVPR'2020)	-	53.09	31.75	-
VSLNet*(ACL'2020)	70.46	54.19	35.22	50.02
$ACRM\_Sub\_GS$*	69.89	50.46	31.13	48.45
$ACRM\_Sub\_TH$*	72.07	56.91	36.56	51.39
$ACRM\_Dot\_GS$*	72.15	**57.93**	37.15	52.60
$ACRM\_Dot\_TH$*	**73.47**	57.53	**38.33**	**53.01**

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND THE STATE-OF-THE-ARTS ON TACOS DATASET. THE SYMBOL '*' MEANS THE I3D FEATURES ARE ADOPTED.

Method	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	mIoU
CTRL(ICCV'2017)	18.32	13.30	-	-
SM-RL(CVPR'2019)	20.25	15.95	-	-
ABLR(AAAI'2019)	19.50	9.40	-	13.40
SAP(AAAI'2019)	-	18.24	-	-
ACL(WACV'2019)	24.17	20.01	-	-
MLVI(AAAI'2019)	20.15	15.23	-	-
TripNet(CVPR'2019)	23.95	19.17	-	-
CBP(AAAI'2020)	27.31	24.79	19.10	21.59
GDP(AAAI'2020)	24.14	-	-	16.18
2D-TAN(AAAI'2020)	37.29	25.32	-	-
SCDM(NIPS'2019)	26.11	21.17	-	-
ExCL*(EMNLP'2019)	45.50	28.00	13.80	-
DRN*(CVPR'2020)	-	23.17	-	-
VSLNet*(ACL'2020)	29.61	24.27	20.03	24.11
$ACRM\_Sub\_GS$*	49.29	**39.34**	26.12	36.44
$ACRM\_Sub\_TH$*	51.09	38.37	25.82	36.59
$ACRM\_Dot\_GS$*	**51.26**	38.27	26.59	37.31
$ACRM\_Dot\_TH$*	51.19	38.79	**26.94**	**37.42**

- SAP [3]: This method integrates the semantic concepts of the queries into the moment candidate generation to obtain discriminative candidates.
- ACL [10]: This method extracts the semantic concepts from verb-obj pairs in the queries and encodes visual concepts in the video to enhance the localization.
- MLVI [15]: This multilevel language and vision integration model generates the query-specific moment candidates by incorporating the query feature to the R-C3D model [13].
- TripNet [16]: This RL-based model utilizes the state processing module to encode the cross-modal features with gated-attention.
- CBP [25]: This model proposed to predict the boundaries based on semantic cues and aggregate contextual information through the self-attention mechanism.
- GDP [19]: The model employs a graph convolutional to capture relationships between the multi-level semantics generated by a frame feature pyramid.
- 2D-TAN [26]: This model employs a two-dimensional temporal map to capture the temporal relations and learn more discriminative semantics of video moments
- SCDM [47]: This model employs a semantic conditioned dynamic modulation mechanism, which employs sentence semantics to modulate the temporal convolution process for better correlating the sentence related video contents.
- ExCL [18]: This model predicts the frame indices of temporal boundaries from the concatenated frame features and the query feature.
- DRN [48]: This method regresses the temporal distances to the boundary frames of the segment from each frame, and uses a regression model to improve the interaction between the predicted and the ground truth location.
- VSLNet [20]: This method proposes a video span localizing network based on the standard span-based Question-Answering (QA) framework, and employs a query-guided highlighting strategy for prediction.

2) Performance Analysis

The performance of four variants of our proposed model (ACRM) are reported and analyzed. The variants are based on the different combination of interaction modeling function (i.e., element-wise multiplication or subtraction) and normalization method (i.e., **tanh** or Gauss). Dot, Sub, TH, and GS are used to represent element-wise multiplication, subtraction, **tanh** activation, and Gauss distribution normalization, respectively. For example, $ACRM\_Dot\_GS$ denotes our model adopts Gauss distribution normalization and element-wise multiplication.

The results of different approaches on the Charades-STA dataset and TACoS dataset are reported in Table I and Table II, respectively. The best performance is highlighted in bold and the best results of the compared baselines are underlined.

From the results, we can have the following observations. For the Charades-STA dataset, the proposed ACRM models outperform all the competitors by a large margin in all metrics except for $ACRM\_Sub\_GS$, which fails to beat SCDM and VSLNet with a comparable result. Compared to ExCL which uses the concatenation of the cross-modal features, the proposed models achieve around 8%, 13%, 16%, and 3% absolute improvements in terms of different metrics.

For the TACoS dataset, the four variants of the proposed model achieve the new state-of-the-art performance in terms of all metrics. Particularly, the proposed $ACRM$ model outperforms the best baseline ExCL with 6% and 10% improvements on the "R@1, IoU=0.3" and "R@1, IoU=0.5" metrics, respectively. It verifies the benefits of exploiting the cross-modal interactions and employing the internal-frame predictor to reinforce the localization process. Moreover, it is worth noting that our model makes an even larger improvement over VSLNet in the more challenging metrics "R@1, IoU=0.7" and "mIoU" by 6% and 13%, demonstrating the superiority of our

TABLE III
ABLATION STUDIES OF THE PROPOSED MODEL ON CHARADES-STA AND TACOS DATASETS WHERE ATT, IFP DENOTE THE ATTENTION MODULE AND THE INTERNAL-FRAME PREDICTOR, RESPECTIVELY. THE "✓" SYMBOL MARKS A COMPONENT IS ENABLED.

Method	ATT	IFP	Charades-STA				TACoS			
			R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	mIoU	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	mIoU
ExCL (EMNLP'2019)			65.10	44.10	22.60	-	45.50	28.00	13.80	-
$ACRM\_Dot\_GS_m$			71.45	55.27	35.91	51.31	48.16	37.39	24.42	35.08
$ACRM\_Dot\_TH_m$			71.96	56.16	37.63	52.13	48.46	37.44	25.52	35.33
$ACRM\_Dot\_GS_a$	✓		71.99	56.13	36.18	51.63	49.32	38.34	26.49	36.43
$ACRM\_Dot\_TH_a$	✓		72.72	57.34	37.07	52.49	49.71	37.94	26.54	36.66
$ACRM\_Dot\_GS$	✓	✓	72.15	**57.93**	37.15	52.60	**51.26**	38.27	26.59	37.31
$ACRM\_Dot\_TH$	✓	✓	**73.47**	57.53	**38.33**	**53.01**	51.19	**38.79**	**26.94**	**37.42**

model. The good performance of our model is attributed to the combining effects of the interaction modeling, the attention module for important words, and the utilization of the internal frames. In the next section, we analyze the contribution of different modules by ablation studies.

Overall, all the variants of our method outperform the baselines consistently across all cases. Comparing the performance of the four variants, we can see that the $ACRM\_Dot\_TH$, achieves the best performance over both datasets. Besides, using element-wise multiplication can achieve substantial improvement over the use of subtractions, indicating that the element-wise multiplication is more effective in modeling the cross-modal feature interactions.

C. Ablation Study

We conduct an ablation study to examine the effectiveness of all the modules in our model, including different cross-modal interaction methods, the attention module, and the internal frame predictor. The results of the ablation study are reported in Table III. ExCL is used as the baseline here. For the internal-frame predictor, the influence of different trade-off hyper-parameter λ is analyzed. We set the ExCL as the baseline where the last hidden state of the query feature and the video frame features are concatenated. Because of space limitations, we only demonstrate the element-wise multiplication models ($ACRM\_Dot\_TH$ and $ACRM\_Dot\_GS$), and the enabled components are marked with a "✓" symbol in Table III.

1) Effects of the Cross-modal Component

$ACRM\_Dot\_TH_m$ and $ACRM\_Dot\_GS_m$ have not used the attention module and the internal frame predictor. Comparing to ExCL, the main difference is that ExCL uses concatenation to fuse the video frame feature and query feature, and the above two methods employ the element-wise multiplication to model the interactions between the video frame feature and mean-pooled query features. From the comparison results on the two datasets illustrated in Table III, we can observe that both models outperform ExCL consistently on the TACoS dataset in terms of all metrics and also achieves significant improvement on the Charades-STA dataset. Specifically, $ACRM\_Dot\_GS_m$ surpasses ExCL by 6.3%, 11.1%, 12.3% on three metrics, respectively. It validates the importance of modeling the interactions between video features and query features in video moment retrieval instead of a simple concatenation. Particularly, the feature concatenation only maintains the unimodal cues of two modalities, which fails to obtain a stable representation. On the contrary, the proposed cross-modal processing module can fully capture the inter-modal interactions and exploit more reliable cross-modal information than the feature concatenation.

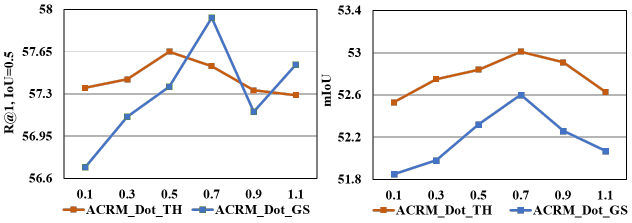


Fig. 3. The R@1, IoU=0.5 and mIoU performance of the proposed ACRM_Dot_TH and ACRM_Dot_GS model with the different parameter λ on the Charades-STA dataset.

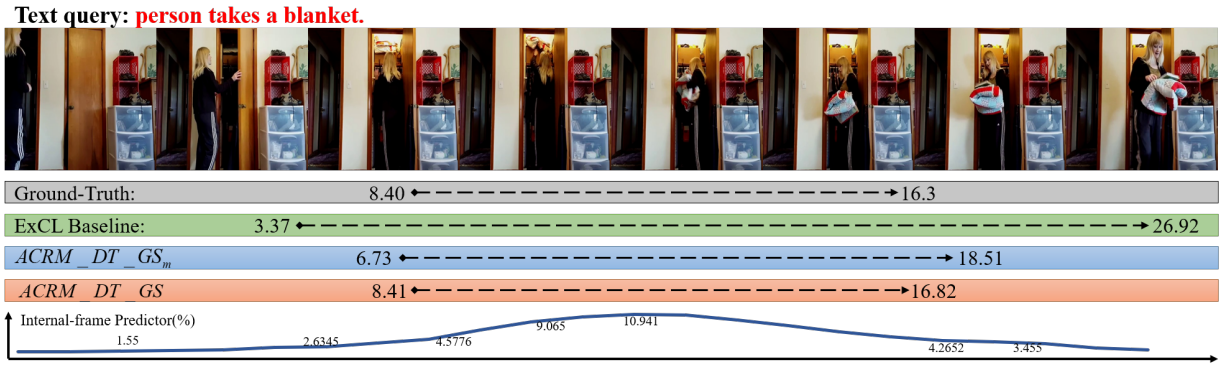
2) Effects of the Attention Component

Compared to the model without attention (i.e., the first two methods in the table), the performance can be consistently improved over all metrics when the attention mechanism is employed ($ACRM\_Dot\_TH_a$ and $ACRM\_Dot\_GS_a$). It demonstrates that the mean-pooling strategy is inadequate to exploit the correlations over the query contexts because the contributions of each word to the query representation are assumed to be equal. Therefore, the incorporation of the attention module is crucial to highlight the contribution of important words in the query and thus enhances the interactions between video frames and the corresponding query.

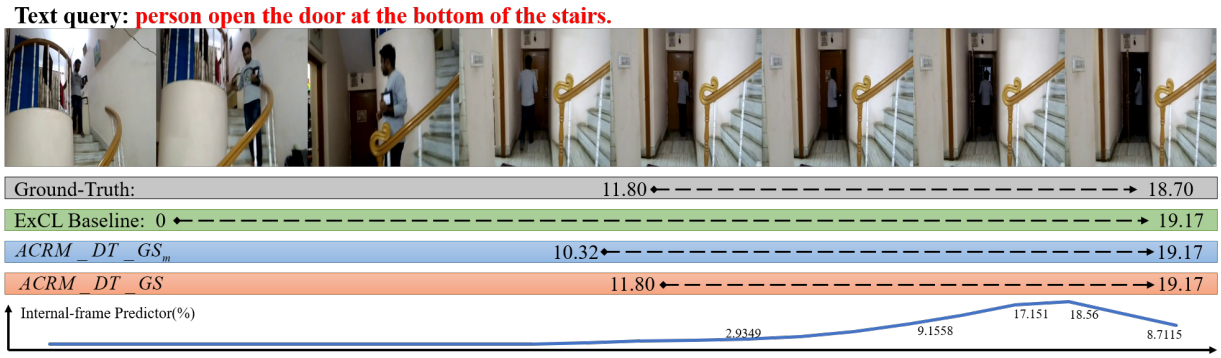
3) Effects of the Internal-frame Predictor

The effects of the internal-frame predictor are analyzed on the Charades-STA dataset through tuning the trade-off parameter λ in the loss function. As illustrated in Figure 3, the performance of our ACRM model with internal-frame predictor outperforms the ones which have not used the internal-frame predictor ($ACRM\_Dot\_TH_a$ and $ACRM\_Dot\_GS_a$). This demonstrates that it is important to consider the internal frames in the modeling, which also contains useful information for boundary detection. In addition, it also indicates that although the integration of an additional internal frame predictor is simple, it is very effective to leverage the internal frame information.

Moreover, we could also observe that when λ increases from 0.1 to 1.1, the reported results of "mIoU" metric only vary in the scope of 1%, indicating the robustness of incorporate the



(a) The moment retrieval result of the proposed ACRM model



(b) The moment retrieval result of the proposed ACRM model

Fig. 4. Moment retrieval results of different models performed on the Charades-STA dataset. All the above examples are the R@1 results. The internal-frame prediction scores are obtained through $ACRM\_Dot\_GS$ model.

internal-frame predictor in the model. With the increase of λ, the variation of the performance follows a general trend, i.e., rises at first and then starts to decline. The optimal value of λ is 0.7, where all the methods obtain the best performance or the competing results on both datasets. It is expected because excessive information of the internal frames may cover the information of the boundaries frames, and thus hurt the accuracy of boundary detection. On the other side, notice that we only incorporate the prediction score of internal frames during the training process. When λ is set to a large value, the parameters of the model will be fine-tuned to predict the internal frames which will not be taken into consideration during the evaluation, resulting in performance degradation.

D. Qualitative Results

1) Visualization Results of Moment Retrieval

In this section, some qualitative examples of the ACRM model and the baseline ExCL model for moment retrieval are illustrated in Figure 4. In the figure, the internal-frame prediction scores are also provided.

Figure 4(a) presents an instance with a relatively simple query. Obviously, ExCL is incapable of returning the required moment. Instead, it returns the entire process of a woman opening her closet and leaving with a blanket because the ExCL model only maintains the information of each modality by concatenating cross-modal features, and hence fails to model the interaction between two modalities. However, the interaction is quite crucial in this video, because there are many frames with similar semantic scenes outside the desired moment, which confuse the model, resulting in an inaccurate prediction of the end time.

In contrast, our $ACRM\_Dot\_GS_m$ model without the attention module and the internal-frame predictor can already achieve a relatively good result with "R@1, IoU=67.06%", which is attributed to the use of adequate interaction modeling of two modalities. It can effectively identify the clips with high correlation and excludes those with low relevance to determine the temporal boundaries. By incorporating the attention module and the internal-frame predictor, our full model localizes the desired moment with a high accuracy of "R@1, IoU=93.71%". Specifically, the query attention module emphasizes the keyword "take" which greatly enhances the prediction of the start time. As for the end time point, we also observe that the internal-frame prediction scores of the inner frames are 3-5 times higher than that of the outsiders, which indicates that the inner prediction component refines the boundaries by implicitly forcing the model to abandon the similar but irrelevant outside frames.

In addition, Figure 4(b) shows an example of a complex query. Similarly, ExCL returns the entire video, which is somehow unreasonable since the object door does not appear in the video at the beginning. ExCL is confused about these frames with low correlation and fails to identify the frames with higher relevance. On the contrary, our

Fig. 5. Visualization results of the frame-by-word attention. The darker the color is, the larger the related attention value is. "GT" represents the ground truth boundaries.

$ACRM\_Dot\_GS_m$ model which only replaces the feature concatenation with an interaction function, successfully excludes frames with low correlation and can return relatively good temporal coordinates. It again demonstrates the importance of considering the cross-modal interactions. For our full model $ACRM\_Dot\_GS$, the precise temporal boundaries are obtained, owing to the refinement of the internal-frame predictor and the attention module.

2) Visualization Results of Frame-by-word Attentions

To verify the effectiveness of the attention module, the qualitative attention weights of a video-query pair is illustrated in Figure 5, where the darker the color is, the larger its represented attention weight is. As in the figure, the attention weight of the word 'person' is always very large since this concept is the major object appearing across all frames. In contrast, for the words 'blanket' and 'laughing', our model assigns much larger attention weights in the first three frames than that in the last two frames where the concept 'blanket' does not appear and the person stops 'laughing'. Finally, some words like 'to', 'be' and 'the' are very small across all frames since these words provide little information for the moment localization.

V. Conclusion

In this paper, we presented an attentive cross-modal relevance matching model (ACRM) to retrieve the relevant moment in an untrimmed video given a specific query. Different from previous methods using a simple concatenation for boundary detection, we highlight the importance of modeling the interactions between the video frame features and query features. In addition, an attention module is integrated into the model to capture more accurate frame-query relations. Moreover, our model exploits the information in the internal frame to enhance the model learning process for boundary prediction by incorporating an internal-frame predictor in the objective function. Extensive experiments have been performed on two benchmark datasets to evaluate the proposed model. Experimental results show that our model substantially outperforms several state-of-the-art baselines by a large margin. Additional ablation studies also validate the effectiveness of each module in our model.

Acknowledgment

This work is supported by the National Natural Science Foundation of China, No.: 61902223, No.: 61573273; The Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014; Young creative team in universities of Shandong Province, No.2020KJN012; Jinan 20 projects in universities, No.2018GXRC0.

References

[1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

[3] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8199–8206.

[4] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang, "Multi-level policy and reward-based deep reinforcement learning framework for image captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1372–1383, 2019.

[5] W. Zhu, X. Wang, and W. Gao, "Multimedia intelligence: When multimedia meets artificial intelligence," *IEEE Transactions on Multimedia*, 2020.

[6] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[7] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: spatial-temporal attention mechanism for video captioning," *IEEE transactions on multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[8] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *2018 ACM Multimedia Conference*. ACM, 2018, pp. 843–851.

[9] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 2018, pp. 15–24.

[10] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 245–253.

[11] B. Jiang, X. Huang, C. Yang, and J. Yuan, "Cross-modal video moment retrieval with spatial and language-temporal attention," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 217–225.

[12] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 162–171.

[13] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5783–5792.

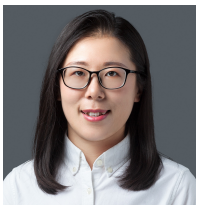
[14] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Text-to-clip video retrieval with early fusion and re-captioning," *arXiv preprint arXiv:1804.05113*, vol. 2, no. 6, p. 7, 2018.

[15] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9062–9069.

[16] M. Hahn, A. Kadav, J. M. Rehg, and H. P. Graf, "Tripping through time: Efficient localization of activities in videos," *arXiv preprint arXiv:1904.09936*, 2019.
[17] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," *arXiv preprint arXiv:2001.06680*, 2020.
[18] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann, "ExCL: Extractive Clip Localization Using Natural Language Descriptions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1984–1990. [Online]. Available: https://www.aclweb.org/anthology/N19-1198
[19] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, and X. Li, "Rethinking the bottom-up framework for query-based video localization." in *AAAI*, 2020, pp. 10 551–10 558.
[20] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 6543–6554. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.585
[21] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9159–9166.
[22] C. Rodriguez, E. Marrese-Taylor, F. S. Saleh, H. LI, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
[23] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *European conference on computer vision*. Springer, 2012, pp. 144–157.
[24] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257.
[25] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction." in *AAAI*, 2020, pp. 12 168–12 175.
[26] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," *arXiv preprint arXiv:1912.03590*, 2019.
[27] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3628–3636.
[28] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942–1950.
[29] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 4325–4334.
[30] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *The European Conference on Computer Vision (ECCV)*, September 2018.
[31] F. Feng, L. Nie, X. Wang, R. Hong, and T.-S. Chua, "Computational social indicators: a case study of chinese university ranking," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 455–464.
[32] F. Feng, X. He, Y. Liu, L. Nie, and T.-S. Chua, "Learning on partial-order hypergraphs," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1523–1532.
[33] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid, "Weakly-supervised alignment of video with text," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4462–4470.
[34] D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Visual semantic search: Retrieving videos via complex textual queries," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2657–2664.
[35] J. Dong, X. Li, and C. G. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.
[36] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, and X. Wang, "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9346–9355.
[37] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 19–27.
[38] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[41] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[42] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, 2015, pp. 3294–3302.
[43] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 510–526.
[44] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010
[45] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4555–4564.
[46] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 334–343.
[47] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *Advances in Neural Information Processing Systems*, 2019, pp. 536–546.
[48] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 287–10 296.

Haoyu Tang received his B.S. degree from Xi'an Jiaotong University, China, in 2016. He is currently an intern scholar in Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, and an Ph.D. candidate in software engineering department, Xi'an Jiaotong University, China. His research interests include machine learning, multimedia retrieval.

Jihua Zhu received the B.E. degree in automation from Central South University, China, and the Ph.D. degree in pattern recognition and intelligence system from Xian Jiaotong University, China, in 2004 and 2011, respectively. He is currently an Associate Professor with the School of Software Engineering, Xian Jiaotong University. His research interests include computer vision and machine learning.



Meng Liu is currently a Professor with the School of Computer Science and Technology, Shandong Jianzhu University. She received the M.S. degree in computational mathematics from Dalian University of Technology, China, in 2016. Her research interests are multimedia computing and information retrieval. Various parts of her work have been published in top forums and journals, such as SIGIR, MM, and IEEE TIP. She has served as reviewers for various conferences and journals, such as ACM MM 2019/2020, AAAI 2020, IEEE TIP, IEEE TKDE, JVCI, and INS.

Zan Gao received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He is currently a Full Professor with the Shandong Artificial Intelligence Institute, Shan dong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. From 2011 to 2018, he worked with the School of Computer Science and Engineering, Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology. From 2009 to 2010, he was a visiting scholar with the School of Computer Science, Carnegie Mellon University, USA, and worked with Prof. A. G. Hauptmann. From July 2016 to January 2017, he worked with Prof. T.-S. Chua with the School of Computing of National University of Singapore as a visiting scholar. His research interests include artificial intelligence, multimedia analysis and retrieval, and machine learning.

Zhiyong Cheng is currently a Professor with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences). He received the Ph.D degree in computer science from Singapore Management University in 2016, and then worked as a Research Fellow in National University of Singapore. His research interests mainly focus on large-scale multimedia content analysis and retrieval. His work has been published in a set of top forums, including ACM SIGIR, MM, WWW, TOIS, IJCAI, TKDE, and TCYB. He has served as the PC member for several top conferences such as SIGIR, MM, IJCAI, AAAI. and the regular reviewer for journals including TKDE, TIP, TMM.