Reasoning Step-by-Step: Temporal Sentence Localization in Videos via Deep Rectification-Modulation Network

Daizong Liu[1], Xiaoye Qu[2], Jianfeng Dong[3], Pan Zhou[1*]
[1]Huazhong University of Science and Technology
[2]Huawei Cloud
[3]Zhejiang Gongshang University
dzliu@hust.edu.cn, quxiaoye@huawei.com,
dongjf24@gmail.com, panzhou@hust.edu.cn

Abstract

Temporal sentence localization in videos aims to ground the best matched segment in an untrimmed video according to a given sentence query. Previous works in this field mainly rely on single-step attentional frameworks to align the temporal boundaries by a soft selection. Although they focus on the visual content relevant to the query, these attention strategies are insufficient to model complex video contents and restrict the higher-level reasoning demand for temporal relation. In this paper, we propose a novel deep rectification-modulation network (RMN), transforming this task into a multi-step reasoning process by repeating rectification and modulation. In each rectification-modulation layer, unlike existing methods directly conducting the cross-modal interaction, we first devise a rectification module to correct implicit attention misalignment which focuses on wrong position during the interaction process. Then, a modulation module is developed to model the frame-to-frame relation with the help of specific sentence information for better correlating and composing the video contents over time. With multiple such layers cascaded in depth, our RMN progressively refines video and query interactions, thus enabling a further precise localization. Experimental evaluations on three public datasets show that the proposed method achieves state-of-the-art performance.

1 Introduction

Localizing activities in videos (Regneri et al., 2013; Yuan et al., 2016; Gavrilyuk et al., 2018; Feng et al., 2018; Feng et al., 2019) is an important topic in information retrieval systems. As most videos contain activities of interest with complicated background contents, these videos cannot be directly indicated by a pre-defined list of action classes. Recently, a new task called temporal sentence localization in videos (Gao et al., 2017; Anne Hendricks et al., 2017) is proposed to tackle this problem, attracting great interests from both vision and language communities (Liu et al., 2020; Qu et al., 2020). Given an untrimmed video, this task aims to infer the start and end timestamps of a target video segment which contains the interested activity according to a given sentence query.

Traditional methods (Gao et al., 2017; Liu et al., 2018; Ge et al., 2019; Chen and Jiang, 2019; Anne Hendricks et al., 2017) are based on sliding windows, which first sample candidate video segments and then compare the sentence with each video segment separately to calculate the matching relationships. These methods cannot achieve precise alignment between video and sentence, thus leading to inaccurate temporal boundaries. Recently, some works (Chen et al., 2018; Chen et al., 2019; Zhang et al., 2019b; Zhang et al., 2019a) try to avoid this problem by designing end-to-end models. They first integrate the features of the whole video with sentence information and then utilize LSTM or CNN layer to compose such integrated video features for further segment localization. Although these methods achieve promising results, there are still some problems need to be concerned.

First, previous works formally adopt single-step attention for multi-modal feature interaction, which limits the modeling power for two reasons: 1) It can not mine sufficient relationship between modalities; 2) Once the cross-modal relation focuses on the wrong position without further calibration, it can heavily jeopardize the localization performance. For example, as shown in Figure 1 (left), the video expresses

 ∗Corresponding Author: Pan Zhou.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1841–1851
Barcelona, Spain (Online), December 8-13, 2020

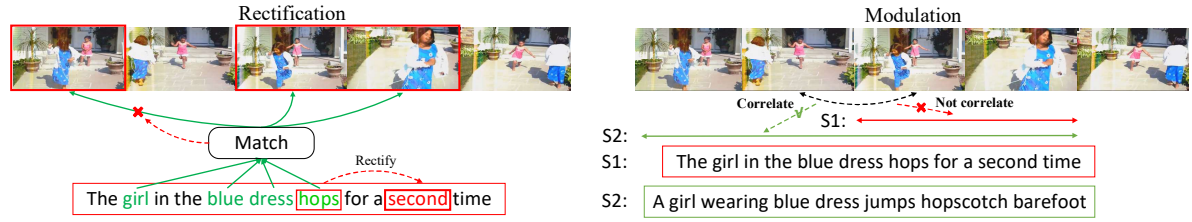

Figure 1: Illustration of rectification and modulation modules for precisely localization. **Left:** The rectification module helps correct the attention to focus on the best matched position. **Right:** The modulation module correlates the video contents with different weights referring to different sentence semantics.

the activity "the girl in the blue dress hops for a second time". All frames contain the similar visual appearance with "girl" and "blue dress", and single-step reasoning may guide the model focus on the action word "hops". It is hard to lead the model directly pay more attention on the adjective "second". Therefore, a multi-steps reasoning framework needs to be developed for not only rectifying the attention errors from the previous reasoning step, but also helping model gradually focus on the most matched words or frames. Second, previous works mainly focus on aligning the sentence information with video clips. Although it is crucial to capture such cross-modal relation between two modalities for highlighting the matched words or frames, the self-relation among video frames is also important for correlating and composing the sentence related video contents over time. To effectively model temporal activities, such self-relations need modulation by the information from other modality, namely the relations between visual frames should be weighted differently according to different sentence queries. As shown in Figure 1 (right), the video contains multiple segment-sentence annotation pairs, the third frame should be correlated with the second one when querying sentence S2 but this correlation should not be established when given the sentence S1. Therefore, how to modulate the temporal relation among video frames conditioned on the matched words from the whole sentence is vital for this task.

In this paper, we propose a novel rectification-modulation network (RMN), which modulates conditioned temporal relation with multiple reasoning steps for temporal sentence localization in videos. In the rectification module, to avoid the error accumulation of the wrong relation from previous reasoning step, we adopt the initial modal feature as a global information flow to correct the attention errors. In the modulation module, we modulate the temporal relation among frames according to the sentence semantics for better correlating sentence-related video contents over time. With multiple such rectification-modulation layers cascaded in depth, our model can reasoning higher-order multi-modal interaction step-by-step, providing more accurate video segment boundaries.

In summary, this paper makes following contributions:

- We propose a novel rectification-modulation network (RMN), which adopts a multi-step reasoning framework to gradually capturing higher-order multi-modal interaction.
- The rectification module utilizes initial multi-modal features as the global information to help our model rectify the attention which focuses on the wrong position from the previous reasoning step.
- The modulation module considers the self-modal relation between video frames conditioned on the sentence semantics. In this way, each frame can be associated with most matched words for correlating the interested video contents.
- We conduct experiments on three public datasets, and verify the effectiveness of our proposed RMN with the superiority over the state-of-the-art methods.

2 Related Work

Temporal sentence localization in videos is a new task introduced recently (Gao et al., 2017; Anne Hendricks et al., 2017), which aims to localize the most relevant video segment from a video with text descriptions. Traditional methods (Liu et al., 2018; Gao et al., 2017) adopt a two-stage multi-modal matching strategy which firstly sample candidate segments from a video, and subsequently integrate query with segment representations via a matrix operation. However, these methods lack a comprehensively structure for effective multi-modal features interaction. Based on such multi-modal matching

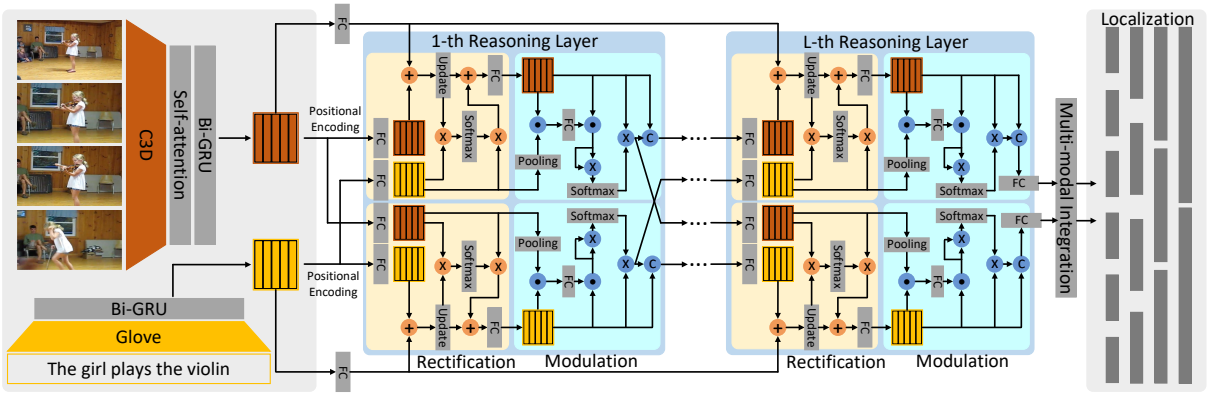

Figure 2: An overview of our proposed rectification-modulation network (RMN). We first embed both visual and language representations by the multi-modal encoders. Then, multi-step rectification-modulation layers are developed to correlate and compose the video contents referring to the sentence-related information. At last, we integrate the multi-modal features for moment localization.

framework, some works (Xu et al., 2019; Chen and Jiang, 2019; Ge et al., 2019) integrate the sentence representation with those video segments individually, and then evaluated their matching relationships through the integrated features. For instance, Xu (Xu et al., 2019) introduce a multi-level model to integrate visual and textual features and further re-generate queries as an auxiliary task. Ge (Ge et al., 2019) and Chen (Chen et al., 2018) capture the evolving fine-grained frame-by-word interactions between video and query to enhance the video representation understanding.

Recently, other works (Chen et al., 2018; Wang et al., 2020; Zhang et al., 2019b; Zhang et al., 2019a; Yuan et al., 2019a; Mithun et al., 2019) propose to directly integrate sentence information with each fine-grained video clip unit, and predict the temporal boundary of the target segment by gradually merging the fusion feature sequence over time. Wang (Wang et al., 2020) aggregate contextual information by explicitly modeling the relationship between the current element and its neighbors. Zhang (Zhang et al., 2019a) model relations among candidate segments with the guidance of the query information. To modulate temporal convolution operations, Yuan (Yuan et al., 2019a) and Mithun (Mithun et al., 2019) introduce the sentence information as a critical prior to compose and correlate video contents.

Although existing methods perform well in this task, all of them adopt a single-step model and only consider aligning the sentence information with video clips, ignoring to associate the video frames conditioned on the sentence features for more precisely moment localization. In this paper, we develop a modulation module to modulate the conditioned temporal relation for contents correlating and composing. Moreover, we repeat the rectification-modulation layer multiple times for deeper reasoning.

3 The Proposed RMN Model

Given an untrimmed video $\boldsymbol{V}$ and a sentence query $\boldsymbol{Q}$, the task aims to determine the start and end timestamps (s, e) of a specific video segment, which corresponds to the activity of the given sentence query. Formally, we represent the video as $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{T}$ frame-by-frame, and denote the given sentence query as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N}$ word-by-word. With the training set $\{\boldsymbol{V}, \boldsymbol{Q}, (s, e)\}$, in this paper, we propose a deep rectification-modulation network (RMN) to learn to predict the most relevant video segment boundary $(\hat{s}, \hat{e})$. As shown in Figure 2, our method contains four parts: multi-modal encoding, multi-step rectification-modulation layers, multi-modal integration, and moment localization.

3.1 Video and Query Encoding

For video encoding, we first extract the frame-wise features by a pre-trained C3D network (Tran et al., 2015), and then employ a self-attention (Vaswani et al., 2017) module to learn the semantic dependencies in the long video context. Considering the sequential characteristic in video, a bi-directional GRU (Chung et al., 2014) is further utilized to incorporate the contextual information. For query information encoding, we first extract the word embeddings by Glove (Pennington et al., 2014), and then feed them

into another bi-directional GRU to integrate the sequential information. We denote the embedddings of video and query as $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times d}$ and $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N} \in \mathbb{R}^{N\times d}$, respectively.

3.2 Multi-Step Rectification-Modulation Layers

In the task of temporal sentence localization in videos, besides understanding the video clip contents, how to capture their temporal correlations plays an even more important role. Luckily, the query sentence presents rich semantic indications on such important correlations, providing crucial information to temporally associate and compose the consecutive video contents over time. Based on the above considerations, we propose a modulation module, which modulates the temporal frame-to-frame relation conditioned on the sentence semantic information for better composing the video contents. To avoid error interaction focused on the wrong position, we additionally develop a rectification module to correct the attention error from previous reasoning step. We conduct rectification and modulation on both video and query to enhance the information flows. With multiple such layers cascaded in depth, our model can reasoning higher-order multi-modal interaction for more precise video segment localization.

Notations. At l-th rectification-modulation layer, we define the multi-modal representation inputs and outputs as $\hat{\boldsymbol{V}}^{l-1}, \hat{\boldsymbol{Q}}^{l-1}$ and $\hat{\boldsymbol{V}}^{l}, \hat{\boldsymbol{Q}}^{l}$, respectively. We also denote the multi-modal hidden states from previous reasoning layer as $\boldsymbol{H}_V^{l-1}, \boldsymbol{H}_Q^{l-1}$, which are utilized as constraints for cross-modal interaction and conditions for self-modal interaction. Specifically, we initialize $\hat{\boldsymbol{V}}^{0}, \hat{\boldsymbol{Q}}^{0} = \boldsymbol{V} + \mathrm{PE}(\boldsymbol{V}), \boldsymbol{Q} + \mathrm{PE}(\boldsymbol{Q})$ with positional encoding (Vaswani et al., 2017) which takes additional positional knowledge to enhance the semantic information, and we set the initial hidden states as $\boldsymbol{H}_V^{0}, \boldsymbol{H}_Q^{0} = \hat{\boldsymbol{V}}^{0}, \hat{\boldsymbol{Q}}^{0}$, respectively.

Rectification Module. Given the multi-modal representations and hidden states from the previous layer, we first aim to rectify the attention error if the learned relation from previous reasoning step is focused on the wrong position. Specifically, we utilize the initial modal features $\boldsymbol{V}, \boldsymbol{Q}$ as global information to regularize and re-correct the the multi-modal flow $\hat{\boldsymbol{V}}^{l-1}, \hat{\boldsymbol{Q}}^{l-1}$ by an update gate:

$$\boldsymbol{Z}_V^{l} = \sigma(\boldsymbol{W}_Z \hat{\boldsymbol{V}}^{l-1} + \boldsymbol{W}_v \boldsymbol{V}), \quad \boldsymbol{Z}_Q^{l} = \sigma(\boldsymbol{W}_z \hat{\boldsymbol{Q}}^{l-1} + \boldsymbol{W}_q \boldsymbol{Q}), \tag{1}$$

$$\tilde{\boldsymbol{V}}^{l-1} = (\boldsymbol{1} - \boldsymbol{Z}_V^{l}) \odot \boldsymbol{V} + \boldsymbol{Z}_V^{l} \odot \hat{\boldsymbol{V}}^{l-1}, \quad \tilde{\boldsymbol{Q}}^{l-1} = (\boldsymbol{1} - \boldsymbol{Z}_Q^{l}) \odot \boldsymbol{Q} + \boldsymbol{Z}_Q^{l} \odot \hat{\boldsymbol{Q}}^{l-1}, \tag{2}$$

where σ is sigmoid function, $\boldsymbol{W}_Z, \boldsymbol{W}_z, \boldsymbol{W}_v, \boldsymbol{W}_q$ are the parameters of linear layers. $\odot$ denotes the element-wise multiplication. With such rectified representations of two modalities, we further utilize cross information flow from other modality to enhance the current modal representation for each modality. Instead of directly computing the cross-relation between representations $(\tilde{\boldsymbol{V}}^{l-1}, \tilde{\boldsymbol{Q}}^{l-1})$, we consider more detailed latent clues from the hidden states $(\boldsymbol{H}_Q^{l-1}, \boldsymbol{H}_V^{l-1})$ from previous reasoning step which can provide more discriminative information for each modality. Following the co-attention mechanism (Lu et al., 2016), we calculate the correlation matrix of cross-modal instances as follows:

$$\boldsymbol{M}_V^{l} = (\boldsymbol{W}_V \tilde{\boldsymbol{V}}^{l-1})(\boldsymbol{W}_H \boldsymbol{H}_Q^{l-1})^T, \quad \boldsymbol{M}_Q^{l} = (\boldsymbol{W}_Q \tilde{\boldsymbol{Q}}^{l-1})(\boldsymbol{W}_h \boldsymbol{H}_V^{l-1})^T, \tag{3}$$

where $\boldsymbol{W}_V, \boldsymbol{W}_Q, \boldsymbol{W}_H, \boldsymbol{W}_h$ are the learnable parameters. Each row of $\boldsymbol{M}_V^{l}$ denotes the similarity of all word features to a specific frame feature, and each row of $\boldsymbol{M}_Q^{l}$ represents the similarity of all frame features to a specific word feature. The value of each similarity will be high if the word-frame pair is relevant or it will be low. To aggregate cross-modal information $\boldsymbol{I}_V^{l}, \boldsymbol{I}_Q^{l}$ for $\tilde{\boldsymbol{V}}^{l-1}, \tilde{\boldsymbol{Q}}^{l-1}$, we utilize a weighted summation strategy based on the correlation matrix $\boldsymbol{M}_V^{l}, \boldsymbol{M}_Q^{l}$ as follows:

$$\boldsymbol{I}_V^{l} = \mathrm{Softmax}(\boldsymbol{M}_V^{l})(\boldsymbol{W}_H \boldsymbol{H}_Q^{l-1}), \quad \boldsymbol{I}_Q^{l} = \mathrm{Softmax}(\boldsymbol{M}_Q^{l})(\boldsymbol{W}_h \boldsymbol{H}_V^{l-1}). \tag{4}$$

Therefore, we can get the enhanced rectified video features $\boldsymbol{S}_V^{l}$ and enhanced rectified sentence features $\boldsymbol{S}_Q^{l}$ by a simple addition function like (Fukui et al., 2016) on two information flows by:

$$\boldsymbol{S}_V^{l} = \tanh(\boldsymbol{I}_V^{l} + \boldsymbol{W}_V \tilde{\boldsymbol{V}}^{l-1}), \quad \boldsymbol{S}_Q^{l} = \tanh(\boldsymbol{I}_Q^{l} + \boldsymbol{W}_Q \tilde{\boldsymbol{Q}}^{l-1}). \tag{5}$$

Modulation Module. After obtaining the enhanced rectified video and sentence features, it is also important to capture their temporal correlations among each modality. Like the cross-modal attention

mechanism, we can directly calculate the frame-to-frame or word-to-word relations and compute the normalized weights $\boldsymbol{A}$ for each instance in each modality by:

$$\boldsymbol{A}_V^l = \text{Softmax}((\boldsymbol{W}_S\boldsymbol{S}_V^l)(\boldsymbol{W}_S\boldsymbol{S}_V^l)^T), \quad \boldsymbol{A}_Q^l = \text{Softmax}((\boldsymbol{W}_s\boldsymbol{S}_Q^l)(\boldsymbol{W}_s\boldsymbol{S}_Q^l)^T), \tag{6}$$

where $\boldsymbol{W}_S, \boldsymbol{W}_s$ are the parameters of linear layers. Although such naive self-attention matrix $\boldsymbol{A}$ estimates the frame-to-frame and word-to-word importance, the relations which can only be identified conditioned on information from the other modality can not be captured. For example, if the video contains multiple moment-sentence annotation pairs, the relations between different visual frames should be weighted differently according to different given sentence query. As the given sentence query presents rich semantic indications on such important correlations for better correlating and composing the consecutive video contents over time, we tend to modulate the temporal frame-to-frame relations referring to the sentence semantics for improving the self-relation matrix $\boldsymbol{A}$ in Eq. (6) by:

$$\boldsymbol{C}_V^l = \sigma(\text{MeanPool}(\boldsymbol{H}_Q^{l-1})) \otimes \boldsymbol{e}_T, \quad \boldsymbol{C}_Q^l = \sigma(\text{MeanPool}(\boldsymbol{H}_V^{l-1})) \otimes \boldsymbol{e}_N, \tag{7}$$

$$\hat{\boldsymbol{S}}_V^l = \sigma(\boldsymbol{W}_1(\boldsymbol{W}_2\boldsymbol{S}_V^l \odot \boldsymbol{W}_3\boldsymbol{C}_V^l)) \odot \boldsymbol{S}_V^l, \quad \hat{\boldsymbol{S}}_Q^l = \sigma(\boldsymbol{W}_4(\boldsymbol{W}_5\boldsymbol{S}_Q^l \odot \boldsymbol{W}_6\boldsymbol{C}_Q^l)) \odot \boldsymbol{S}_Q^l, \tag{8}$$

$$\boldsymbol{A}_V^l = \text{Softmax}((\boldsymbol{W}_S\hat{\boldsymbol{S}}_V^l)(\boldsymbol{W}_S\hat{\boldsymbol{S}}_V^l)^T), \quad \boldsymbol{A}_Q^l = \text{Softmax}((\boldsymbol{W}_s\hat{\boldsymbol{S}}_Q^l)(\boldsymbol{W}_s\hat{\boldsymbol{S}}_Q^l)^T), \tag{9}$$

where $(\cdot \otimes \boldsymbol{e}_T)$ is the outer product to produce a matrix by repeating the vector on the left for T times, $W_{\{1,2,3,4,5,6\}}$ are the parameters of linear layers. By expanding the sentence/video features after mean pooling, the conditional information $\boldsymbol{C}$ from other modalities can be acquired. Channels of feature $\boldsymbol{S}$ would be further activated or deactivated by such channel-wise gates condition $\boldsymbol{C}$, which shares the similar spirit with Squeeze and Excitation Network (Hu et al., 2018) and the Gated Convolution (Gehring et al., 2017). Therefore each temporal feature map can absorb the sentence semantic information, and further activate the self-correlation matrix for better associating and composing the sentence-related video contents. Words can also enhance its contextual meaning in the same way. We apply matrix multiplication on self-relation weights and multi-modal features to generate self-interacted information by:

$$\boldsymbol{H}_V^l = \boldsymbol{A}_V^l(\boldsymbol{W}_S\boldsymbol{S}_V^l), \quad \boldsymbol{H}_Q^l = \boldsymbol{A}_Q^l(\boldsymbol{W}_s\boldsymbol{S}_Q^l), \tag{10}$$

where we denote such self-interacted information as the hidden states $\boldsymbol{H}_V^l, \boldsymbol{H}_Q^l$ for the input of next reasoning layer. We concatenate the self-interacted information with the enhanced rectified features as the final output of current rectification-modulation layer:

$$\hat{\boldsymbol{V}}^l = \text{Concat}([\boldsymbol{S}_V^l, \boldsymbol{H}_V^l]), \quad \hat{\boldsymbol{Q}}^l = \text{Concat}([\boldsymbol{S}_Q^l, \boldsymbol{H}_Q^l]). \tag{11}$$

3.3 Multi-modal Integration

After multiple rectification-modulation layers, we utilize two linear layers on the two-modal outputs and then get the final video/sentence representations $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{Q}}$. We additionally utilize a cosine similarity function (Mithun et al., 2019) to transfer the dimension of $\hat{\boldsymbol{Q}}$ as the same as $\hat{\boldsymbol{V}}$. To further emphasize crucial contents and weaken inessential parts among each modality, we design a gate function as follow:

$$\boldsymbol{g}_V = \sigma(\boldsymbol{W}_G\hat{\boldsymbol{V}} + \boldsymbol{b}_G), \quad \boldsymbol{g}_Q = \sigma(\boldsymbol{W}_g\hat{\boldsymbol{Q}} + \boldsymbol{b}_g), \tag{12}$$

where $\boldsymbol{W}_G, \boldsymbol{W}_g$ and $\boldsymbol{b}_G, \boldsymbol{b}_g$ are learnable parameters. We then integrate the multi-modal features by:

$$\boldsymbol{f} = \text{Concat}([\boldsymbol{g}_V \odot \hat{\boldsymbol{V}}, \boldsymbol{g}_Q \odot \hat{\boldsymbol{Q}}]), \quad \boldsymbol{f} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_T\}. \tag{13}$$

3.4 Moment Localization

With the integrated representation $\boldsymbol{f}$, we further apply a bi-directional GRU network to absorb more contextual evidences in temporal domain. To predict the target video segment, we first pre-define a set of candidate moments $\Phi_t = \{(\hat{s}_{t,i}, \hat{e}_{t,i})\}_{i=1}^{N_\Phi}$ with multi-scale windows (Yuan et al., 2019a) at each time t, where N_Φ is the number of moments at current time-step. Then, we adopt a Conv1d layer to score

these candidate moments and predict corresponding offsets $\hat{\delta}_t = \{(\hat{\delta}^s_{t,i}, \hat{\delta}^e_{t,i})\}_{i=1}^{N_\Phi}$ of them relative to the ground-truth. The confidence scores $cs_t = \{cs_{t,i}\}_{i=1}^{N_\Phi}$ for these moments can be formulated as follows:

$$cs_{t,i} = \sigma(\text{Conv1d}(\boldsymbol{f}_t)) \in (0, 1), \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function to normalize the confidence scores. Also, the temporal offsets of each candidate moment i at time t can be predicted by another Conv1d layer:

$$(\hat{\delta}^s_{t,i}, \hat{\delta}^e_{t,i}) = \text{Conv1d}(\boldsymbol{f}_t). \tag{15}$$

Therefore, the final predicted moment i of time t can be presented as $(\hat{s}_{t,i} + \hat{\delta}^s_{t,i}, \hat{e}_{t,i} + \hat{\delta}^e_{t,i})$.

Training. To learn the confidence scoring rule for candidate moments, we compute the IoU (Intersection over Union) score $IoU_{t,i}$ between each candidate moment $(\hat{s}_{t,i}, \hat{e}_{t,i})$ with the ground truth (s_t, e_t). We adopt the alignment loss function to train the scoring rule as follows:

$$\mathcal{L}_{align} = -\frac{1}{TN_\Phi} \sum_{t=1}^{T} \sum_{i=1}^{N_\Phi} IoU_{t,i} log(cs_{t,i}) + (1 - IoU_{t,i}) log(1 - cs_{t,i}). \tag{16}$$

Since parts of the pre-defined candidates are coarse in boundaries, to learning to offsets prediction, we only need to fine-tune the localization offsets of positive moment samples. We treat the candidate moment as a positive sample if its $IoU_{t,i}$ is larger than an IoU threshold τ. The moment boundary loss for offsets prediction can be formulated as:

$$\mathcal{L}_b = \frac{1}{N_{pos}} \sum_{j}^{N_{pos}} \mathcal{R}_1(\hat{\delta}^s_j - \delta^s_j) + \mathcal{R}_1(\hat{\delta}^e_j - \delta^e_j), \tag{17}$$

where N_{pos} denotes the number of positive moments, and $\mathcal{R}_1$ is the smooth L1 loss. Both two losses are jointly considered for training with the balanced hyper-parameter α as:

$$\mathcal{L} = \mathcal{L}_{align} + \alpha \mathcal{L}_b. \tag{18}$$

Inference. We first rank all candidate moments according to their predicted confidence scores, and then adopt a non-maximum suppression (NMS) to select "Top n" moments as the prediction.

4 Experiments

4.1 Datasets and Evaluation

Activity Caption (Krishna et al., 2017): It contains 20k untrimmed videos with 100k descriptions from more complicated human activities in daily life. Since the test split is withheld for competition, following public split, we 37,417, 17,505, and 17,031 query-segment pairs for training, validaiton and testing.

Charades-STA (Sigurdsson et al., 2016): It focuses on indoor activities where the videos are 30 seconds on average. There are 12408 and 3720 moment-query pairs in the training and testing sets respectively.

TACoS (Regneri et al., 2013): This dataset is collected from cooking scenarios which contains 127 videos. We use the same split as (Gao et al., 2017), which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing.

Evaluation Metrics. Following previous works (Gao et al., 2017; Yuan et al., 2019a), we adopt "R@n, IoU=m" as our evaluation metrics. The "R@n, IoU=m" is defined as the percentage of at least one of top-n selected moments having IoU larger than m.

4.2 Implementation Details

We utilize the 112×112 pixels shape of every frame of videos as input, and apply C3D (Tran et al., 2015) for ActivityNet Caption and TACoS, I3D (Carreira and Zisserman, 2017) for Charades-STA to encode the videos. We set the length of video feature sequences to 200 for Activity Caption and TACoS,

Method	Activity Caption						Charades-STA			
	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
MCN	39.35	21.36	6.43	68.12	53.23	29.70	17.46	8.01	48.22	26.73
TGN	45.51	28.47	-	57.32	43.33	-	-	-	-	-
CTRL	47.43	29.01	10.34	75.32	59.17	37.54	23.63	8.89	58.92	29.57
ACRN	49.70	31.67	11.25	76.50	60.34	38.57	20.26	7.64	71.99	27.79
QSPN	52.13	33.26	13.43	77.72	62.39	40.78	35.60	15.80	79.40	45.40
CBP	54.30	35.76	17.80	77.63	65.89	46.20	36.80	18.87	70.94	50.19
SCDM	54.80	36.75	19.86	77.29	64.99	41.53	54.44	33.43	74.43	58.08
ABLR	55.67	36.79	-	-	-	-	-	-	-	-
GDP	56.17	39.27	-	-	-	-	39.47	18.49	-	-
CMIN	63.61	43.40	23.88	80.54	67.95	50.73	-	-	-	-
Ours	**67.01**	**47.41**	**27.21**	**87.03**	**75.64**	**56.76**	**59.13**	**36.98**	**87.51**	**61.02**

Table 1: Performance comparisons on the Activity Caption and Charades-STA dataset.

Method	R@1,IoU=0.1	R@1,IoU=0.3	R@1,IoU=0.5	R@5,IoU=0.1	R@5,IoU=0.3	R@5,IoU=0.5
MCN	3.11	1.64	1.25	3.11	2.03	1.25
CTRL	24.32	18.32	13.30	48.73	36.69	25.42
ABLR	34.70	19.50	9.40	-	-	-
ACRN	24.22	19.52	14.62	47.42	34.97	24.88
QSPN	25.31	20.15	15.23	53.21	36.72	25.30
TGN	41.87	21.77	18.90	53.40	39.06	31.02
GDP	39.68	24.14	13.50	-	-	-
CMIN	32.48	24.64	18.05	62.13	38.46	27.02
SCDM	-	26.11	21.17	-	40.16	32.18
CBP	-	27.31	24.79	-	43.64	37.40
Ours	**42.17**	**32.21**	**25.61**	**68.75**	**54.20**	**40.58**

Table 2: Performance comparisons on the TACoS dataset.

64 for Charades-STA. As for sentence encoding, we utilize Glove word2vec (Pennington et al., 2014) to embed each word to 300 dimension features. The hidden state dimension of BiGRU networks is set to 512. During moment localization, we adopt convolution kernel size of [16, 32, 64, 96, 128, 160, 192] for Activity Caption, [8, 16, 32, 64] for TACoS, and [16, 24, 32, 40] for Charades-STA. We set the stride of them as 0.5, 0.125, 0.125. We then set the high-score threshold τ to 0.45, and the balance hyper-parameter α to 0.001 for Activity Caption, 0.005 for TACoS and Charades-STA. We adopt 5 rectification-modulation layers for all datasets. We train our model with an Adam optimizer with leaning rate $8 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}$ for Activity Caption, TACoS, and Charades-STA respectively.

4.3 Compared Methods

We compare our proposed model with the state-of-the-art baseline methods, which can be divided into two classes: 1) Sliding window based methods: **MCN** (Anne Hendricks et al., 2017), **CTRL** (Gao et al., 2017), and **ACRN** (Liu et al., 2018). 2) Cross-modal interaction based single-step methods: **TGN** (Chen et al., 2018), **QSPN** (Xu et al., 2019), **CBP** (Wang et al., 2020), **SCDM** (Yuan et al., 2019a), **ABLR** (Yuan et al., 2019b), **GDP** (Chen et al., 2020), and **CMIN** (Zhang et al., 2019b).

4.4 Performance Comparison

The performance comparisons of existing state-of-the-art methods on three datasets are shown in Table 1 and Table 2. We can observe that the our RMN achieves a new state-of-the-art performance under all evaluation metrics and benchmarks, demonstrating the superiority of our proposed model. For localizing complex human activities in Activity Caption and Charades-STA datasets, our model surpasses others with clear margin on both R@1 and R@5 metrics. Specifically, our method brings 3.33% and 3.55% absolute improvements in the strict metrics "R@1, IoU=0.7", and brings 6.03% and 2.94% absolute improvements in the strict metrics "R@5, IoU=0.7" on two datasets, respectively. For TACoS where the cooking activities take place in the same kitchen scene with some slightly varied cooking objects, it is hard to localize such fine-grained activities. However, our model still achieve the best performance on both R@1 and R@5 metrics with a clear margin.

The main reasons for our proposed RMN outperforming the state-of-the-art methods lies in two folds.

Components	Setting	R@1,IoU=0.1	R@1,IoU=0.3	R@1,IoU=0.5	R@5,IoU=0.1	R@5,IoU=0.3	R@5,IoU=0.5
Baseline	1 layer	38.95	28.73	22.39	65.60	52.19	36.78
Interaction Layers	w/o R&M	34.31	25.19	18.92	60.85	47.83	32.27
	w/o REC	36.34	27.42	20.43	64.01	50.68	35.47
	w/o MOD	36.26	26.70	20.37	63.81	50.91	34.55
Number of Interaction Layers	3 layers	41.09	30.53	24.07	67.74	53.05	38.13
	5 layers	**42.17**	32.21	**25.61**	**68.75**	**54.20**	40.58
	7 layers	41.78	**32.59**	25.44	68.51	54.16	**40.86**

Table 3: Ablation study of the rectification-modulation interaction layer on TACoS dataset.

Components	Setting	R@1,IoU=0.1	R@1,IoU=0.3	R@1,IoU=0.5	R@5,IoU=0.1	R@5,IoU=0.3	R@5,IoU=0.5
Rectification Module	w/ ADD	**42.17**	**32.21**	**25.61**	**68.75**	**54.20**	**40.58**
	w/o ADD	39.13	29.47	23.62	66.98	52.25	37.87
	w/ CONC	40.87	31.08	24.35	67.94	53.43	39.17
	w/ MEM	40.48	30.30	24.12	67.51	51.87	39.11
Modulation Module	w/ FMUL	**42.17**	**32.21**	**25.61**	**68.75**	**54.20**	**40.58**
	w/o FMUL	40.01	30.19	22.68	66.47	52.33	38.16
	w/ MUL	40.76	30.90	23.36	67.15	52.29	38.89
	w/ FC	40.93	30.83	24.11	67.64	52.62	38.91
	w/ CROG	41.44	32.03	24.80	68.24	53.41	39.56

Table 4: Ablation studies of the rectification module and modulation module on the TACoS dataset.

First, instead of only capturing the cross-modal relations (eg. SCDM, GDP), we additionally modulate the temporal relations among frames referring to sentence-related semantic information. Such modulation module helps model better correlate and compose the most relevant video contents according to the sentence over time. Second, compared to single-step interaction methods (eg. CMIN, TGN), our multi-step reasoning process can gradually focus on the most contributed frames and words for better interaction. Also, rectification module is able to correct the attention error from previous reasoning step.

4.5 Ablation Study

How does rectification-modulation interaction layer help? The proposed rectification-modulation interaction layer is the key to our method to reason more higher-level interaction between two modalities. As shown in Table 3, we set the number of such interaction layer to 1 as our baseline model. Here, we first investigate the ablation study on such interaction layer with three variants of models: **w/o R&M** (without using both rectification and modulation), **w/o REC** (only without using rectification) and **w/o MOD** (only without using modulation). We can find that **w/o R&M** achieves the worst performance as it lacks of efficient interaction. Both **w/o REC** and **w/o MOD** achieve relatively higher results but still lower than the result of the default setting, which indicates that both rectification and modulation are crucial for this task. Moreover, we also investigate the influence of the number of stacked interaction layers. As shown in Table 3, we find that more layers can improve the performance thanks to our rectification module, and our model achieves the best result with 5 layers.

How does rectification module help? The rectification module integrates the previous reasoning output with a global information flow from initial modal features. We conduct the ablation study on the usage of such global flow with different settings: we denote **w/ ADD** as the addition operation illustrated in Eq. (1); we remove the global flow as **w/o ADD**; we replace the addition operation with concatenation (**w/ CONC**); and we utilize all previous layer features (Nam et al., 2017) including the initial feature as the global flow (**w/ MEM**). As shown in Table 4, the **w/o ADD** achieves the worst performance as it lacks of attention rectification. The **w/ ADD** outperforms than the other two models, it denotes that the initial modal features are more effective than all previous layers features for rectifying the attention error.

How does modulation module help? To evaluate the contribution of our conditional modulation module, we conduct an ablation study on different condition methods. **w/ FMUL**: our proposed channel-wise condition method in Eq. (8); **w/o FMUL**: we capture self-relation without conditional information in Eq. (6); **w/ MUL**: we replace the FMUL with directly element-wise multiplication on $\boldsymbol{C}$ and $\boldsymbol{S}$; **w/ FC**: we replace the FMUL by using FC layer to fuse each temporal feature unit with each sentence representation; **w/ CROG**: In stead of using FMUL, we utilize a cross-gate (Feng et al., 2018) as condition. As shown in Table 4, we can find that **w/ FMUL** performs the best.

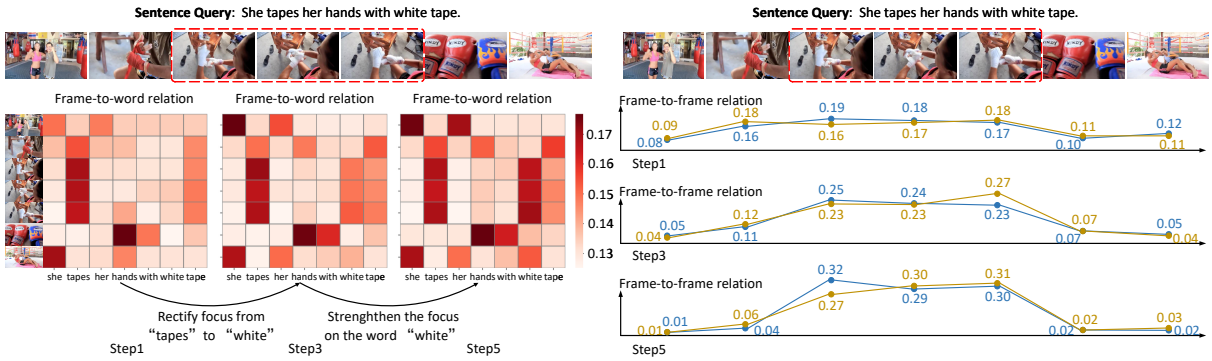

Figure 3: Visualization on the frame-to-word and frame-to-frame relations of different reasoning steps.

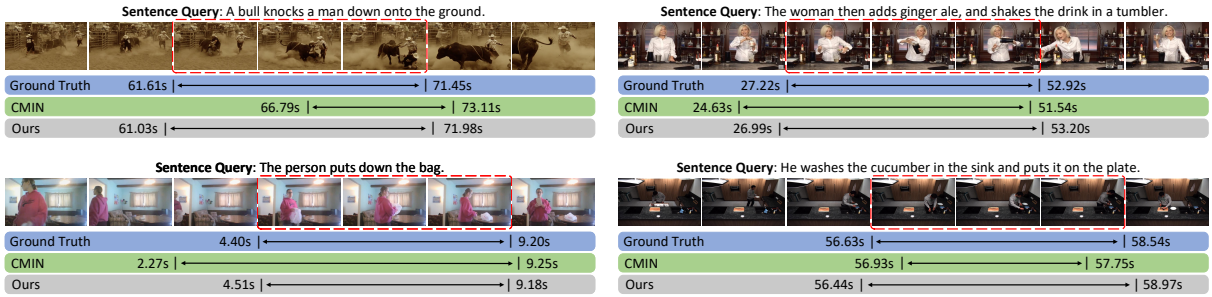

Figure 4: Qualitative results sampled from Activity Caption (top), Charades-STA (down left) and TACoS (down right) datasets, respectively.

4.6 Qualitative Results

To investigate how our rectification and modulation modules work step-by-step, we show one visualization example on Activity Caption dataset in Figure 3. As shown in the left part, we first visualize the frame-to-word relation learned from the rectification module in different reasoning steps. At the first step, 2-5th frames have similar word-related attention which focus on the same words "tapes" and "tape". Although the 2th frame has the similar visual appearance like the 3-5th frames, the people tapes the red tape, not the mentioned "white" tape. With the step goes on, the rectification module adjusts the attention of previous step from "tape" to "white", leading to distinguish the 2th frame from the 3-5th. At step 5, the frame-to-word relations are more distinguishable and the attention on the target frames is focused more on the word "white". It demonstrates that our rectification helps model rectify the attention weights for better grounding the segment boundaries. In the right part, we visualize the attention weights on frame-to-frame relation utilizing softmax function. Similar to the frame-to-word relation, in the first step, the 2th frame is taken as a noisy frame which disturbs the frame-wise correlating. Thanks to the rectification module, with the reasoning step goes on, the weight of noisy frame is getting smaller and our modulation module can better capture the temporal relation referring to the matched words. To qualitatively validate the effectiveness of our method, we also show some qualitative examples from three datasets in Figure 4, where our model provides more precisely video segment boundaries.

5 Conclusion

In this paper, we propose a deep multi-step rectification-modulation network (RMN) for temporal sentence localization in videos. Different from previous single-step methods, we utilize the initial multi-modal features as global information flows to correct the attention errors from previous reasoning step in the rectification module. In the modulation module, we modulate the temporal relation among video frames referring to sentence semantics for better associating and composing sentence-related video contents over time. With multiple such rectification-modulation layers cascaded in depth, our model can reasoning the matched video segment according to the selected words from the given sentence query step-by-step. Extensive experiments on three real-world datasets validate the effectiveness of our method.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812.

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

Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NIPS)*.

Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. 2018. Video re-localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Spatio-temporal video re-localization by warp lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1288–1297.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5958–5966.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning (ICML.*

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pages 7132–7141.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715.

Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 15–24.

Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross- and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4070–4078.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 289–297.

Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11592–11601.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 299–307.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4280–4288.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pages 510–526.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069.

Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kassim. 2016. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102.

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019a. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 534–544.

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019b. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166.

Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019a. 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 (CVPR)*, pages 1247–1257.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 655–664.