Context-aware Biaffine Localizing Network for Temporal Sentence Grounding

Daizong Liu[1*] Xiaoye Qu[2*] Jianfeng Dong[3] Pan Zhou[1†] Yu Cheng[4] Wei Wei[1]
Zichuan Xu[5] Yulai Xie[1]
[1]Huazhong University of Science and Technology [2]Huawei Cloud
[3]Zhejiang Gongshang University [4]Microsoft AI
[5]Dalian University of Technology
{dzliu, panzhou, weiw, ylxie}@hust.edu.cn quxiaoye@huawei.com
dongjf24@gmail.com yu.cheng@microsoft.com z.xu@dlut.edu.cn

Abstract

This paper addresses the problem of temporal sentence grounding (TSG), which aims to identify the temporal boundary of a specific segment from an untrimmed video by a sentence query. Previous works either compare pre-defined candidate segments with the query and select the best one by ranking, or directly regress the boundary timestamps of the target segment. In this paper, we propose a novel localization framework that scores all pairs of start and end indices within the video simultaneously with a biaffine mechanism. In particular, we present a ***C****ontext-aware* ***B****iaffine* ***L****ocalizing* ***N****etwork (CBLN) which incorporates both local and global contexts into features of each start/end position for biaffine-based localization. The local contexts from the adjacent frames help distinguish the visually similar appearance, and the global contexts from the entire video contribute to reasoning the temporal relation. Besides, we also develop a multi-modal self-attention module to provide fine-grained query-guided video representation for this biaffine strategy. Extensive experiments show that our CBLN significantly outperforms state-of-the-arts on three public datasets (ActivityNet Captions, TACoS, and Charades-STA), demonstrating the effectiveness of the proposed localization framework. The code is available at https://github.com/liudaizong/CBLN.*

1. Introduction

Video understanding is a fundamental task in computer vision and has drawn increasing attention over the last years due to its various applications in video event detection [8], video summarization [36, 9, 24], video captioning [18, 6, 20] and temporal action localization [33, 51], etc. Recently, temporal sentence grounding (TSG) [14, 1] has been proposed as an important yet challenging task. This task requires automatically determining the start and end timestamps of a target segment in an untrimmed video that contains an activity semantically corresponding to a given sentence description, as shown in Figure 1. It is substantially more challenging as it needs to not only model the complex multi-modal interactions among vision and language features, but also capture complicated context information for their semantics alignment.

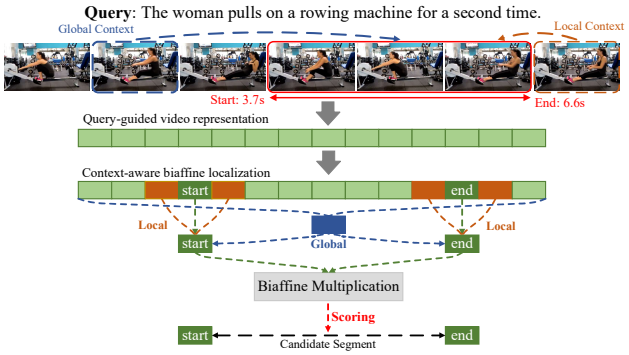

Figure 1. An illustrative example of temporal sentence grounding task. From a new perspective, we propose a biaffine-based localization method that scores all pairs of start and end frames simultaneously with both local and global contexts.

Most previous methods [1, 14, 15, 25, 4, 50, 26, 45] tackle TSG task following a multi-modal matching architecture, which generates multiple candidate proposals of different time intervals and ranks them according to their similarities with the sentence query. These methods severely rely on the quality of proposals, and break the intrinsic temporal structure and global context of videos. Recently, several works [31, 46, 5, 41, 16, 28, 47] directly regress the temporal locations of the target segment. Specifically, they either regress the start/end timestamps based on the entire video representation [46, 28], or predict at each frame to determine whether this frame is a start or end boundary [31, 47, 5]. However, in these methods, start and end features are never jointly considered. Given a video with two target segments that have the same starting action (*open the*

*Equal contributions. †Corresponding author.

door) but end with different actions (*go in*, *go out*), predicting the start point independently may lead to timestamp confusion. Moreover, the ending point also tends to be inaccurate if it is predicted conditioned on the wrong start time.

Different from the aforementioned frameworks, we address the TSG from a new perspective: we reformulate this task by scoring all pairs of start and end indices simultaneously with a biaffine mechanism which interacts characteristics of each pair of possible start and end frames. This biaffine-based architecture is inspired by the dependency parsing task [12] in natural language processing, in which the system predicts a dependency head for each child token and assigns a relation to the head-child pairs. However, there are two main obstacles when the biaffine mechanism used in dependency parsing is applied to TSG task. First, different from adjacent words in a sentence which carry different meaning, video is continuous and the adjacent frames naturally contain visually similar appearance. As shown in Figure 1, the adjacent frames near the segment boundary possess the similar semantics on "woman" and "pulls". Thus, it is difficult to distinguish the specific boundary from adjacent frames without referring to these adjacent features as local context. Second, causalities between word pairs are usually indirect and can be far apart, but in videos, events in different intervals are directly correlated and rely on the whole contents to reason the precise semantics. For example, in Figure 1, without perceiving "a second time" from a global perspective, the first and second time of "pull on a rowing machine" possess similar semantic boundaries but totally different temporal indices. Therefore, the causal relations between the video events which act as global context are essential for understanding the segment.

Based on the above considerations, we propose a novel **C**ontext-aware **B**iaffine **L**ocalizing **N**etwork (CBLN), for temporal sentence grounding. Specifically, we develop a multi-context biaffine localization (MCBL) module which aggregates both local and global contexts to enrich the information of each frame representation. For each frame, we span the entire video features with different window sizes to get multi-scale video events as global contexts, and extract different numbers of adjacent frame features as multi-scale local contexts. The multi-scale local and global contexts are then inter-modulated to produce more adaptive contexts, which serve as the input representation for further biaffine localization by concatenating with frame-wise feature. At last, we obtain the output scores from the biaffine model to identify the similarities of all possible start-end pairs according to the semantics of sentence query. Besides, to provide fine-grained query-guided video representation for above biaffine localization, we also develop a multi-modal self attention (MMSA) module to sufficiently capture dependencies among video frames under the guidance of the sentence description. By jointly learning the overall model, our CBLN is able to localize query in video effectively.

Our main contributions are summarized as follows:

- From a new perspective, we adopt biaffine mechanism to the TSG task. The biaffine-based architecture simultaneously scores all possible pairs of start and end frames for segment localization. Compared to previous methods, it gets rid of complicated proposal design and interacts both start-end timestamps effectively.
- To alleviate the limitation of the biaffine localization, we further develop a multi-context biaffine localization module which utilizes multi-scale local and global contexts to enrich frame representations.
- We conduct extensive experiments to validate the effectiveness of our proposed CBLN on three datasets (ActivityNet Captions, TACoS, and Charades-STA), and show that it significantly outperforms the state-of-the-arts by a large margin.

2. Related Work

Temporal Sentence Grounding. Temporal sentence grounding (TSG) is a new task introduced recently [14, 1, 11] that aims to retrieve video segments using language queries. The early works [1, 15, 25, 48, 4, 50, 26, 45, 43, 23, 22] employ a multi-modal matching architecture that first generates segment proposals, and then ranks them according to the similarity between proposals and the query to select the best matching one. Some of them [14, 1] propose to apply the sliding windows to generate proposals and subsequently integrate the query with segment representations via a matrix operation. Instead of using the sliding windows, latest works [40, 48, 45, 50] directly integrate sentence information with each fine-grained video clip unit, and predict the scores of candidate segments by gradually merging the fusion feature sequence over time. Although those methods achieve promising performances, they are severely limited by the quality of proposals.

To overcome above drawback, recent works [31, 46, 5, 41, 16, 28, 47] directly regress the temporal locations of the target segment. Yuan *et al.* [46] propose a co-attention based network to regress the start and end boundaries of the target segment. To improve the grounding with dense supervisions, Zeng *et al.* [47] regress the distances from each frame to the target start (end) frame. Chen *et al.* [5] propose a graph based bottom-up framework to capture multi-level semantics and encode the plentiful scene relationships.

Different from aforementioned two types of methods, we give a new solution to address the TSG problem. Specifically, we regard each frame as start or end frame to build all possible candidate segments and score all of them simultaneously with a biaffine mechanism [12]. After that, we obtain the output scores for all segments and choose the best one corresponding to the highest value as the grounding

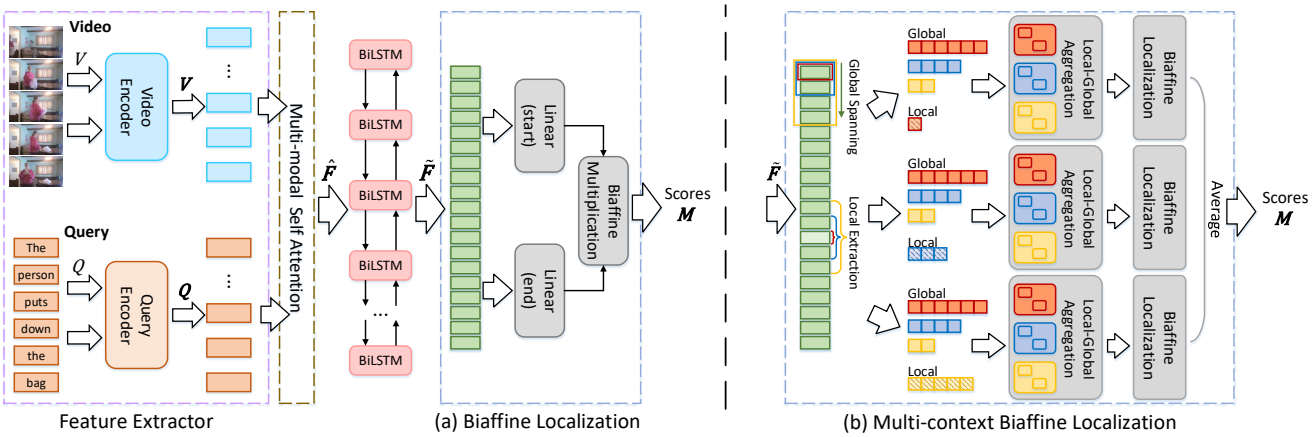

Figure 2. An overview of the proposed architecture for TSG. Given a video and a sentence query, we first encode them by two feature extractors, and further interact them by a multi-modal self attention module. After obtaining query-guided video representations, we exploit a biaffine localization (Figure (a)) to score all possible segments simultaneously. Figure (b) is an improved version of Figure (a) and shows the structure of multi-context biaffine localization module which aggregates multi-scale local and global contexts.

result. Although 2D-TAN [49] also scores all possible segments among the video, it directly max-pools the contained frame features to represent corresponding segment and lose discriminative frame-wise feature. Thus, compared to our detailed cross-modal interaction, this method fails to perform fine-grained interaction. Meanwhile, it captures the proposal-wise relations with convolution layers. Instead, our CBLN incorporates explicit local and global context to capture fine-grained frame-wise relations.

Biaffine based Dependency Parsing. Biaffine mechanism is widely used in dependency parsing [12, 21, 44] which aims to build up a syntactic dependency tree for a given sentence. This task needs to capture all possible relations between the word pairs. Dozat *et al.* [12] are the first work to learn the long-dependency from a head word to a modifier word with a relation label by proposing biaffine. They utilize biaffine operation as a scoring algorithm to determine the syntactic of a phrase from one word to another. Yu *et al.* [44] further adapt biaffine to Named Entity Recognition [13] by reformulating this task as the task of identifying start and end indices, as well as assigning a category to the span by these pairs. By treating input video as text passage, the biaffine mechanism is also applicable to TSG task in principle, because TSG aims to determine whether the segment from one frame to another is a full segment containing the activities described by the sentence query. However, the biaffine is not able to capture the adjacent contexts or correlate the global events, thus may lose local and global details about the scene meaning. In this paper, we enhance the biaffine mechanism by aggregating local-global information.

3. Proposed Method

Given an untrimmed video $\mathcal{V}$ and a sentence query $\mathcal{Q}$, we represent the video as $\mathcal{V} = \{v_t\}_{t=1}^{T}$ frame-by-frame, where v_t is the t-th frame and T is the number of total frames. Similarly, the query with N words is denoted as $\mathcal{Q} = \{q_n\}_{n=1}^{N}$ word-by-word. Temporal sentence grounding (TSG) aims to localize a segment (τ_s, τ_e) starting at timestamp τ_s and ending at timestamp τ_e in video $\mathcal{V}$, which corresponds to the same semantic as query $\mathcal{Q}$.

The key to our Context-aware Biaffine Localizing Network (CBLN) is that we score all pairs of start and end frames simultaneously with a biaffine mechanism by interacting the characteristics of the start-end pairs. As shown in Figure 2, we first utilize two encoders to extract both video and query features, and then introduce a multi-modal self attention (MMSA) to generate the fine-grained video features for localization. Subsequently, we exploit biaffine localization module to score the start and end frame pairs of all possible segments, as shown in Figure 2 (a). In this way, we can get the scores for all candidate segments and choose the best one as the target segment. By enriching the query-guided video representation with local-global contexts, in Figure 2 (b), we further propose a multi-context biaffine localization (MCBL) module for more precise grounding.

3.1. Feature Extractor

Video encoder. For video encoding, we first extract the frame features by a pre-trained C3D network [38], and then add a positional encoding [39] to take positional knowledge. Such position encoding plays a crucial role in distinguishing semantics at diverse temporal locations. Considering the sequential characteristic in videos, a bi-directional GRU [10] is further utilized to incorporate the contextual information along time series. We denote the extracted video features as $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times D}$.

Query encoder. For query encoding, we first extract the word embeddings by the Glove model [29]. We also apply the positional encoding and bi-directional GRU to integrate the sequential information. The final feature of the input sentence query is denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N} \in \mathbb{R}^{N \times D}$.

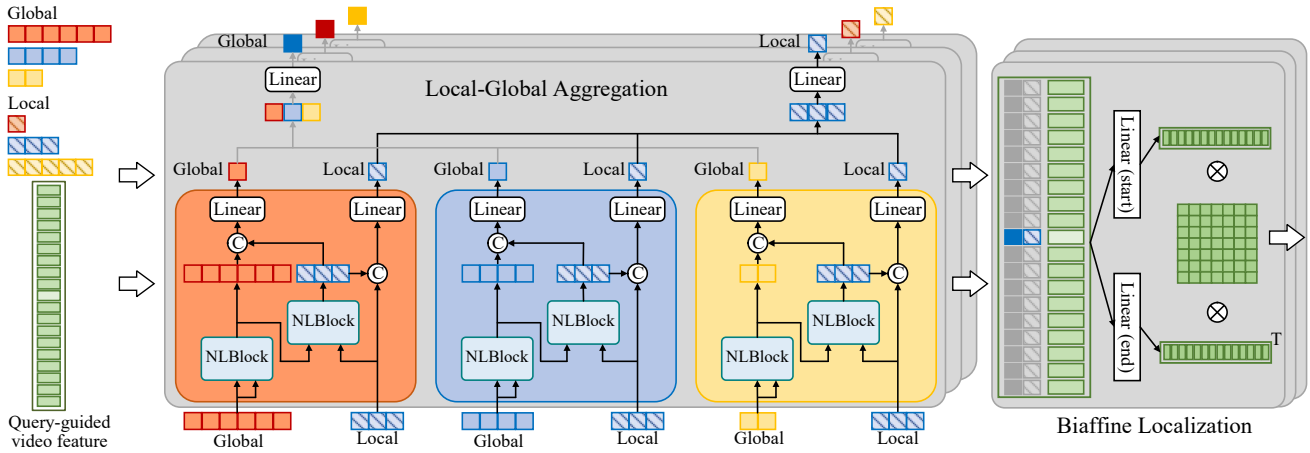

Figure 3. An illustration of the process of aggregating both multi-scale local and global contexts to local-guided global and global-guided local contexts. These aggregated contexts are then concatenated with video representation for biaffine localization.

3.2. Multi-Context Biaffine Localization

3.2.1 Biaffine Localization

Biaffine mechanism is widely used in dependency parsing [12, 21, 44] to assign scores to all possible spans in a sentence, where each span is defined by a pair of start and end words. As TSG can also be reformulated as the task of identifying the start and end frames of a specific segment among a given video, it is appropriate to adapt biaffine mechanism [12] to this task for scoring all possible segment candidates. In our CBLN, before interacting the features of start and end frames using the biaffine mechanism, we first propose a multi-modal self attention (illustrated in Section 3.3) to generate fine-grained query-guided video representation $\widehat{\boldsymbol{F}} = \{\widehat{\boldsymbol{f}}_t\}_{t=1}^{T}$, and then exploit a BiLSTM [32] layer to further aggregate its sequential contexts as:

$$\widetilde{\boldsymbol{F}} = \text{BiLSTM}(\widehat{\boldsymbol{F}}), \tag{1}$$

where $\widetilde{\boldsymbol{F}} = \{\widetilde{\boldsymbol{f}}_t\}_{t=1}^{T}$. Then, as the contexts of the start and end frames are different, we apply two separate linear layers to generate separate hidden representations for each pair of start and end frames:

$$\boldsymbol{h}_p^s = \widetilde{\boldsymbol{f}}_{p_s}\boldsymbol{W}^s + \boldsymbol{b}^s, \tag{2}$$

$$\boldsymbol{h}_p^e = \widetilde{\boldsymbol{f}}_{p_e}\boldsymbol{W}^e + \boldsymbol{b}^e, \tag{3}$$

where p_s and p_e are the start and end indexes of segment p, $\boldsymbol{W}^s, \boldsymbol{W}^e$, $\boldsymbol{b}^s, \boldsymbol{b}^e$ are the learnable parameters of two linear layers. Finally, we employ the biaffine operation over the start and end representations to generate scores $\boldsymbol{M} = \{\boldsymbol{M}_p\}_{p=1}^{T\times T} \in \mathbb{R}^{T\times T}$, which indicates the matching score of all segments. Details can be found in Figure 2 (a). The scoring rule for each segment p can be summarized as:

$$\boldsymbol{M}_p = \sigma(\boldsymbol{h}_p^s \boldsymbol{U}^m (\boldsymbol{h}_p^e)^\top + (\boldsymbol{h}_p^s \oplus \boldsymbol{h}_p^e)\boldsymbol{W}^m + \boldsymbol{b}^m), \tag{4}$$

where $\boldsymbol{U}^m$ and $\boldsymbol{W}^m$ are learnable parameters, $\boldsymbol{b}^m$ is bias, and $\oplus$ denotes element-wise addition, σ is the sigmoid function. After a sigmoid layer, $\boldsymbol{M}_p$ is the score of segment p, which indicates the probability of p matched to the query. Experiments in next section demonstrate that biaffine mechanism achieves a superior performance in TSG task.

3.2.2 Biaffine Localization with Local-Global Contexts

Although biaffine localization module has a strong ability to address TSG, it measures each segment by only considering the features of its start and end frames. As a result, the context of segments can only be drawn from a limited extent of previous BiLSTM layer. To enrich the context information of the start and end frames, we integrate start/end representation with: 1) Local contexts from adjacent frames. Since the adjacent frames near the segment boundary present similar visual appearance to the start/end frames, we aggregate the local contexts to distinguish them for more accurate boundaries grounding. 2) Global contexts from the entire video. We also extract long-range contexts to reason the temporal relations between different events among the video. Moreover, as shown in Figure 2 (b), the multi-scale local and global contexts are further aggregated. Finally, we concatenate each aggregated local-global contexts with corresponding start/end frame features for parallel multi-context biaffine localization.

Local-Global contexts. Given a start/end frame t, we define two types of context features: "local" features $\boldsymbol{R}_t^l$ extracted from adjacent frames, and "global" features $\boldsymbol{R}_t^g$ spanned from the entire video. For "local" features, since the local contexts cover several adjacent frames around the frame t, we directly utilize a window of size K^l on frame t to extract features as follows:

$$\boldsymbol{R}_t^l = \{\widetilde{\boldsymbol{f}}_{t-(K^l/2)}, ..., \widetilde{\boldsymbol{f}}_t, ...\widetilde{\boldsymbol{f}}_{t+(K^l/2)}\}, \tag{5}$$

where $\boldsymbol{R}_t^l \in \mathbb{R}^{K^l\times D}$. To capture more contextual details from local features, we generate local contexts of multiple scales by varying the number of $K^l = \{1, 3, 5\}$.

For "global" features, as the global contexts refer to snippet-level features max-pooled from the long-range

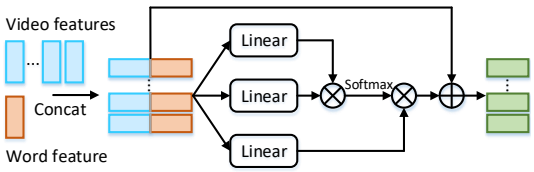

Figure 4. The structure of multi-modal self attention on each word-video feature pair. Since query contains N words, we employ N such modules and average pool the corresponding outputs to generate fine-grained query-guided video representation.

video frames, we define the spanning features $\boldsymbol{R}_t^g$ as a feature bank of snippet features with a window of size K^g as:

$$\boldsymbol{R}_{t,k}^g = \text{maxpool}(\widetilde{\boldsymbol{f}}_{(k-1)T/K^g}, ..., \widetilde{\boldsymbol{f}}_{kT/K^g-1}), \quad (6)$$

$$\boldsymbol{R}_t^g = \{\boldsymbol{R}_{t,1}^g, \boldsymbol{R}_{t,2}^g, ..., \boldsymbol{R}_{t,k}^g, ..., \boldsymbol{R}_{t,(T/K^g)}^g\}, \quad (7)$$

where $\boldsymbol{R}_{t,k}^g \in \mathbb{R}^{1\times D}$ is the max-pooled snippet features across a specific segment, and $\boldsymbol{R}_t^g \in \mathbb{R}^{T/K^g\times D}$ is the global contexts. We also generate multi-scale global contexts by varying $K^g = \{1, 2, 4\}$. We then ensemble these local-global contexts for further aggregation.

Local-Global aggregation. Since both local and global contexts need to be adaptive to the boundary frame at each position, we modulate them to generate local-guided global and global-guided local contexts. In detail, we aggregate all scales of global features with each scale of local feature separately. In Figure 3, we demonstrate the details of multi-scale local-global aggregation. Given global feature of multiple scales and local feature of one specific scale, we first re-weight each local-global pair $(\boldsymbol{R}_t^l, \boldsymbol{R}_t^g)$ by:

$$(\boldsymbol{R}_t^g)' = \text{NLBlock}(\boldsymbol{R}_t^g, \boldsymbol{R}_t^g), \quad (8)$$

$$(\boldsymbol{R}_t^l)' = \text{NLBlock}((\boldsymbol{R}_t^g)', \boldsymbol{R}_t^l), \quad (9)$$

where NLBlock$(\cdot)$ denotes a modified non-local block [42] added with a layer normalization [2] and dropout layer [37]. The first NLBlock is utilized to capture temporal relations between the pooled events. Since feature $(\boldsymbol{R}_t^g)'$ loses specific information on the interested frames, another NLBlock is designed to further model adaptive local contexts by reasoning the adjacent features with global events. At last, we concatenate multi-level global/local contexts and project them through a linear layer to compute the final local-guided/global-guided global/local context for frame t.

Biaffine with multi-context. After receiving the local-guided global and global-guided local contexts from each local-global aggregation module at frame t, we concatenate them to the corresponding frame feature $\widetilde{\boldsymbol{f}}_t$. Based on this feature, we can get a new context-aware query-guided video representation $(\widetilde{\boldsymbol{F}})'$, and then process it by Eq. (2)(3)(4) in biaffine localization module. At last, we average the results of size $T \times T$ from the outputs of multiple biaffine localization modules to produce the final scores $\boldsymbol{M} \in \mathbb{R}^{T\times T}$ for all segments, which is further followed by a sigmoid function. Each value on scores $\boldsymbol{M}$ represents the matching score between a segment with the queried sentence.

3.3. Multi-Modal Self Attention

Building interaction between video and query is a crucial step to provide detailed query-guided video representation for localization. To achieve this goal, recent works [5, 50, 46] interact each word with each frame by a co-attention mechanism. However, these methods only focus on frame-wise cross-modal matching and lack the interaction over long-range video frames under the query guidance, which is essential for consecutive semantics understanding. To learn a better query-guided video representation forming the input of our biaffine-based localization, as depicted in Figure 4, we first concatenate word feature to each frame feature in a video and feed them into a multi-modal self attention module for long-range dependencies capturing. Specifically, given a word q_n, we first construct a joint multi-modal feature by concatenating single word feature $\boldsymbol{q}_n$ to the whole video features $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^T$ as:

$$\boldsymbol{F}_n = \{\boldsymbol{f}_{nt}\}_{t=1}^T, \text{ where } \boldsymbol{f}_{nt} = [\boldsymbol{v}_t; \boldsymbol{q}_n] \in \mathbb{R}^{1\times 2D}. \quad (10)$$

The multi-modal self attention module then takes the multi-modal features $\boldsymbol{F}_n$ as input, and produces a set of query, key and value pair by linear transformations as $\boldsymbol{l}_{nt}^q = \boldsymbol{f}_{nt}\boldsymbol{W}^q$, $\boldsymbol{l}_{nt}^k = \boldsymbol{f}_{nt}\boldsymbol{W}^k$ and $\boldsymbol{l}_{nt}^v = \boldsymbol{f}_{nt}\boldsymbol{W}^v$ at each frame t, where $\boldsymbol{W}^q, \boldsymbol{W}^k, \boldsymbol{W}^v$ are parameters to be learned. We compute the multi-modal self attentive feature $\widehat{\boldsymbol{l}}_{nt}^v$ by:

$$\widehat{\boldsymbol{l}}_{nt}^v = \sum_{t'=1}^{T} \alpha_{(nt,nt')}\boldsymbol{l}_{nt'}^v, \alpha_{(nt,nt')} = \text{Softmax}(\boldsymbol{l}_{nt}^q(\boldsymbol{l}_{nt'}^k)^\top). \quad (11)$$

$\alpha_{(nt,nt')}$ is the weight coefficient computed by a softmax function, and takes into account of the correlation between (n, t) and (n, t') which consists of same word but different frames. Next, we transform $\widehat{\boldsymbol{l}}_{nt}^v$ back to the same dimension as $\boldsymbol{f}_{nt}$ via a linear layer and add it element-wise with $\boldsymbol{f}_{nt}$ to form a residual connection [17]:

$$\widehat{\boldsymbol{f}}_{nt} = \widehat{\boldsymbol{l}}_{nt}^v\boldsymbol{W}^b + \boldsymbol{f}_{nt}. \quad (12)$$

We use $\widehat{\boldsymbol{F}}_n = \{\widehat{\boldsymbol{f}}_{nt}\}_{t=1}^T \in \mathbb{R}^{T\times 2D}$ to denote the collection of $\widehat{\boldsymbol{f}}_{nt}$ at all video frames. Based on $\widehat{\boldsymbol{F}}_n$, we can capture long-range dependencies among the video under the n-th word guidance. In this stage, we utilize N such multi-modal self attention modules to capture different word-video dependencies. The final query-guided video feature representation is average pooled over all N words of the query as:

$$\widehat{\boldsymbol{F}} = \text{Average}(\{\widehat{\boldsymbol{F}}_n\}_{n=1}^N) = \frac{\sum_{n=1}^N \widehat{\boldsymbol{F}}_n}{N}. \quad (13)$$

3.4. Training Details

To train our CBLN, we utilize the scaled Intersection over Union (IoU) values as the supervision signal. Specifically, we compute the IoU score o_p of each segment p with

the ground truth, and scale o_p as the supervision signal by:

$$o_p = o_p / \text{Max}(\boldsymbol{O}), \tag{14}$$

where $\text{Max}(\boldsymbol{O})$ denotes the maximum IoU score among all IoU scores $\boldsymbol{O} = \{o_p\}_{p=1}^{T\times T} \in \mathbb{R}^{T\times T}$. Our network is trained by a binary cross entropy loss as follows:

$$\mathcal{L} = -\frac{1}{T\times T}\sum_{p=1}^{T\times T} o_p \log(\boldsymbol{M}_p) + (1-o_p)\log(1-\boldsymbol{M}_p), \tag{15}$$

where $\boldsymbol{M}_p$ is the score of the segment p in the output $\boldsymbol{M}$.

4. Experiments

4.1. Datasets and Evaluation

ActivityNet Captions. ActivityNet Captions [19] contains 20000 untrimmed videos with 100000 descriptions from YouTube. The videos are 2 minutes on average, and the annotated video clips have much larger variation, ranging from several seconds to over 3 minutes. Following public split, we use 37417, 17505, and 17031 sentence-video pairs for training, validation, and testing respectively.

TACoS. TACoS [30] is widely used on TSG task and contain 127 videos. The videos from TACoS are collected from cooking scenarios, thus lacking the diversity. They are around 7 minutes on average. We use the same split as [14], which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing.

Charades-STA. Charades-STA is built on the Charades dataset [34], which focuses on indoor activities. In total, the video length on the Charades-STA dataset is 30 seconds on average, and there are 12408 and 3720 moment-query pairs in the training and testing sets, respectively.

Evaluation. Following previous works [14, 49, 47], 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

For video encoding, we apply C3D [38] to encode the videos on all three datasets, and also extract the I3D [3] and VGG [35] features on Charades-STA dataset. Since some videos are overlong, we set the length of video feature sequences to 200 for ActivityNet Captions and TACoS datasets, 64 for Charades-STA dataset, respectively. As for sentence encoding, we utilize Glove word2vec [29] to embed each word to 300 dimension features. The hidden state dimensions of bi-directional GRU and BiLSTM are set to 512. We train our model with an Adam optimizer with leaning rate 8×10^{-4}, 3×10^{-4}, 4×10^{-4} for ActivityNet Captions, TACoS, and Charades-STA datasets, respectively. The batch size is set to 64. More model details can be found in our supplementary material.

Table 1. Comparisons on ActivityNet using C3D features.

Method	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
TGN [4]	45.51	28.47	-	57.32	43.33	-
CTRL [14]	47.43	29.01	10.34	75.32	59.17	37.54
ACRN [25]	49.70	31.67	11.25	76.50	60.34	38.57
QSPN [43]	52.13	33.26	13.43	77.72	62.39	40.78
CBP [40]	54.30	35.76	17.80	77.63	65.89	46.20
SCDM [45]	54.80	36.75	19.86	77.29	64.99	41.53
LGI [28]	58.52	41.51	23.07	-	-	-
2D-TAN [49]	59.45	44.51	26.54	85.53	77.13	61.96
CMIN [50]	63.61	43.40	23.88	80.54	67.95	50.73
DRN [47]	-	45.45	24.36	-	77.97	50.30
Ours	**66.34**	**48.12**	**27.60**	**88.91**	**79.32**	**63.41**

Table 2. Comparisons on TACoS using C3D features.

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
ACRN [25]	24.22	19.52	14.62	47.42	34.97	24.88
CTRL [14]	24.32	18.32	13.30	48.73	36.69	25.42
QSPN [43]	25.31	20.15	15.23	53.21	36.72	25.30
CMIN [50]	32.48	24.64	18.05	62.13	38.46	27.02
SCDM [45]	-	26.11	21.17	-	40.16	32.18
CBP [40]	-	27.31	24.79	-	43.64	37.40
TGN [4]	41.87	21.77	18.90	53.40	39.06	31.02
DRN [47]	-	-	23.17	-	-	33.36
2D-TAN [49]	47.59	37.29	25.32	70.31	57.81	45.04
Ours	**49.16**	**38.98**	**27.65**	**73.12**	**59.96**	**46.24**

4.3. Comparisons with state-of-the-arts

Comparisons on ActivityNet Captions. We compare our CBLN with the state-of-the-art methods on the ActivityNet Captions dataset in Table 1. We follow the previous methods to use C3D features for fair comparisons. Particularly, our model outperforms the previously best method DRN [47] by 3.24% and 13.11% absolute improvement in terms of R@1, IoU=0.7 and R@5, IoU=0.7, respectively. Compared to the method 2D-TAN [49], we also outperform them by 6.89%, 3.61%, 1.06%, 3.38%, 2.19% and 1.45% in terms of all metrics, respectively.

Comparisons on TACoS. We compare our CBLN with the state of-the-art methods with the same C3D features in Table 2. On TACoS dataset, the cooking activities take place in the same kitchen scene with slightly varied cooking objects, thus showing the challenging nature of this dataset. Despite its difficulty, our model still reaches highest scores in terms of both R@1 and R@5 when IoU=0.5, and outperforms both 2D-TAN [49] and DRN [47] by a great margin.

Comparisons on Charades-STA. Table 3 reports the grounding results of various methods. Our CBLN reaches the highest results over all evaluation metrics. Specifically, when using the same VGG features, compared to the previously best method 2D-TAN [49], our model brings the absolute improvement of 3.86%, 1.19%, 9.06% and 5.34% on all metrics, respectively. For fair comparisons with GDP [5] and LGI [28], we also perform experiments with same features (i.e., C3D and I3D) reported in their papers. It is ob-

Table 3. Comparisons with state-of-the-arts on Charades-STA.

Method	Feature	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
SAP [7]	VGG	27.42	13.36	66.37	38.15
2D-TAN [49]	VGG	39.81	23.25	79.33	51.15
Ours	VGG	**43.67**	**24.44**	**88.39**	**56.49**
CTRL [14]	C3D	23.63	8.89	58.92	29.57
QSPN [43]	C3D	35.60	15.80	79.40	45.40
CBP [40]	C3D	36.80	18.87	70.94	50.19
GDP [5]	C3D	39.47	18.49	-	-
Ours	C3D	**47.94**	**28.22**	**88.20**	**57.47**
DRN [47]	I3D	53.09	31.75	89.06	60.05
SCDM [45]	I3D	54.44	33.43	74.43	58.08
LGI [28]	I3D	59.46	35.48	-	-
Ours	I3D	**61.13**	**38.22**	**90.33**	**61.69**

Table 4. Ablation studies of the baseline model with different grounding strategies on ActivityNet Captions, where * denote the baseline model with biaffine mechanism.

Model	R@1, IoU=0.7	R@5, IoU=0.7
Baseline (regression)	17.02	45.83
Baseline (proposal-match)	20.10	51.96
Baseline* (biaffine)	**22.74**	**53.79**

vious that our model still performs better. All these results again verify the effectiveness of our model.

4.4. Ablation Studies

In this section, we will perform in-depth ablation studies to evaluate the effect of each component in our CBLN on ActivityNet Captions dataset.

Main ablation studies. First of all, to investigate the effectiveness of biaffine localization module in this paper, we build up three baseline models with different grounding heads. For a fair comparison, we keep video/query encoders and cross-modal attention mechanism [27, 50] consistent in all baselines. Baseline (regression) directly regresses temporal boundary [46, 28], while Baseline (proposal-match) designs pre-defined proposals to match the query [45, 50]. Table 4 shows that the Baseline* (biaffine) with our proposed biaffine mechanism significantly outperforms the other two baselines. It demonstrates that the biaffine mechanism is more suitable for segment localization in TSG task as it can learn detailed start-end frames interaction and get rid of handcrafted proposals compared to previous methods.

Next, to investigate the contribution of the proposed multi-modal self attention (MMSA) module and multi-context biaffine localization (MCBL) module, we also implement three variants of our model as shown in Table 5. Compared to the baseline*, MMSA captures more fine-grained query-video interactions and outperforms it by 1.48% and 3.77% in R@1, and R@5,IoU=0.7, respectively. Moreover, MCBL brings the highest improvement on both two metrics (i.e. 3.74% and 7.19%), which demonstrates its

Table 5. Results of main ablation studies on ActivityNet Captions.

Multi-modal self attention (MMSA)	Multi-context biaffine localization (MCBL)	R@1, IoU=0.7	R@5, IoU=0.7
×	×	22.74	53.79
✓	×	24.22	57.56
×	✓	26.48	60.98
✓	✓	**27.60**	**63.41**

Table 6. Performance comparison with varying different local-global contexts in MCBL module on ActivityNet Captions dataset.

Component	Changes	R@1, IoU=0.7	R@5, IoU=0.7
local extraction	mean-pooling	25.09	59.98
	max-pooling	25.47	61.29
	concatenate	**27.60**	**63.41**
global spanning	sampling	24.88	58.25
	mean-pooling	26.02	61.83
	max-pooling	**27.60**	**63.41**
local scale K^l	{1}	25.45	61.12
	{3}	25.72	61.52
	{5}	25.63	61.49
	{7}	25.17	60.96
	{1,3}	26.76	62.17
	{3,5,7}	27.13	62.63
	{1,3,5}	**27.60**	63.41
	{1,3,5,7}	27.38	**63.49**
global scale K^g	{1}	25.41	61.34
	{2}	25.13	60.99
	{4}	25.06	60.75
	{8}	24.89	60.38
	{1,2}	26.48	62.02
	{2,4,8}	26.91	62.77
	{1,2,4}	27.60	63.41
	{1,2,4,8}	**27.72**	**63.68**

effectiveness of aggregating local-global contexts.

Analysis on local-global contexts generation. As shown in Table 6, we conduct the investigation on the impact of different local-global contexts modeling in multi-context biaffine localization (MCBL) module. First, for local context extraction, we need to preserve the details of each adjacent frame, thus the pooling strategy may lose some discriminative information contained in each frame features. The results also show that the concatenation performs better than both mean- and max-pooling. For global context spanning, it is difficult to work with all the raw features. Therefore, we select representative features by sub-sampling or pooling. In our experiments, we can find that max-pooling is superior to both random sampling and mean-pooling, as random sampling may lose discriminative frame feature and mean-pooling smooths away salient features that are otherwise preserved by max-pooling.

Besides, we also show the influence on different scales of both local K^l and global K^g window sizes. With more kinds of different scales, the model usually performs better than individual scale. For local scale K^l, we choose

Table 7. Performance comparison with varying combinations of modules in local-global aggregation on ActivityNet Captions.

Changes	R@1 IoU=0.7	R@5 IoU=0.7
Full model	**27.60**	**63.41**
remove all non-local blocks	25.05	60.88
replace non-local with linear	25.93	61.74
replace final linear with pooling	26.86	62.78

Table 8. Performance comparison with varying fusion strategies on multiple information on ActivityNet Captions.

Module	Fusion	R@1, IoU=0.7	R@5, IoU=0.7
MMSA	concat+linear	26.84	62.37
	max-pooling	25.53	61.24
	mean-pooling	**27.60**	**63.41**
MCBL	max-pooling	27.46	63.12
	mean-pooling	**27.60**	**63.41**

Table 9. Complexity comparison on ActivityNet. "Speed" denotes the average time to localize one sentence in a given video.

MMSA	MCBL	R@1, IoU=0.7	Speed	Memory (batchsize)
✓	✓	**27.60**	0.18s	10184M (64)
×	✓	26.48	0.14s	8747M (64)
✓	×	24.22	0.09s	5898M (64)
CMIN		23.88	0.08s	5692M (64)
2D-TAN		26.54	0.57s	10572M (16)

$K^g = \{1, 3, 5\}$. As for global scale K^g, the variant with four kinds of scales $\{1, 2, 4, 8\}$ achieves the best result but only performs marginally better than the three-scales one $\{1, 2, 4\}$ at the expense of significantly larger cost of GPU memory. Thus, we choose $K^g = \{1, 2, 4\}$ for global contexts in our experiments.

Analysis on local-global contexts aggregation. To interact information of both local and global contexts, we apply stacked non-local blocks to combine them in a learnable way in the local-global aggregation module of MCBL. As shown in Table 7, we can observe that when removing all non-local blocks and separately passing global and local contexts to latter computation, there is a performance drop of 2.55% and 2.53% in R@1, and R@5, IoU=0.7, respectively. When we replace non-local block with concatenation and a linear layer, there is also a drop of 1.73% and 1.67%. These results demonstrate that the stacked non-local blocks are effective for local-global contexts re-weighting. We also evaluate the performance when we replace the final linear layer (for multiple local/global combination, as shown in Figure 3) with max-pooling, it drops 0.74% and 0.63%.

Analysis on the fusion strategies. To better fuse the multiple information obtained by the outputs of both multi-modal self attention (MMSA) module and multi-context biaffine localization (MCBL) module, we compare different fusion operations as shown in Table 8. We can find that mean-pooling achieves the best performance in both two modules.

Analysis on model Complexity. To investigate the complexity of our model, we give an in-depth study in terms of speed and memory as shown in Table 9. Though our full model is not more efficient than CMIN, it outperforms CMIN with a large margin. Our w/o MCBL method still achieves better performance with similar computational cost (compared to CMIN). Compared to 2D-TAN, our full model performs better and much more efficient.

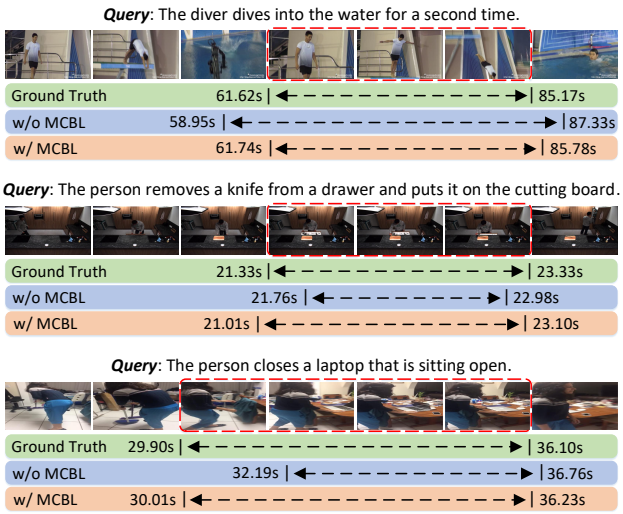

Figure 5. Qualitative results on ActivityNet Captions (top), TACoS (middle), and Charades-STA (bottom) datasets, respectively.

4.5. Qualitative Results

Figure 5 shows some qualitative results from three datasets. Our multi-context biaffine localization module can provide more contextual details about the segment, thus achieves better grounding results.

5. Conclusion

In this paper, we have proposed a novel context-aware biaffine localizing network, called CBLN, for temporal sentence grounding. The key to CBLN is that we reformulate this task from a new perspective for scoring all pairs of start and end indices simultaneously by a biaffine mechanism. To enrich the feature representation of each start/end frame, we additionally integrate them with multi-scale local and global contexts. A multi-modal self attention module is also developed to generate fine-grained query-guided video representation for such biaffine-based strategy. The experiments on three public datasets demonstrate the performance of CBLN, which brings significant improvements over the state-of-the-art methods.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (No.61972448, No.61902347), and the Zhejiang Provincial Natural Science Foundation (No. LQ19F020002).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017. 1, 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

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

[4] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171, 2018. 1, 2, 6

[5] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 1, 2, 5, 6, 7

[6] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–351. Springer, 2020. 1

[7] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8199–8206, 2019. 7

[8] Yu Cheng, Quanfu Fan, Sharath Pankanti, and Alok Choudhary. Temporal sequence modeling for video event detection. In *CVPR*, 2014. 1

[9] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3584–3592, 2015. 1

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 3

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

[12] Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2, 3, 4

[13] Jenny Rose Finkel and Christopher D Manning. Nested named entity recognition. In *Proceedings of theConference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 141–150, 2009. 3

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5275, 2017. 1, 2, 6, 7

[15] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 245–253, 2019. 1, 2

[16] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8393–8400, 2019. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[18] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018. 1

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 6

[20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 1

[21] Ying Li, Zhenghua Li, Min Zhang, Rui Wang, Sheng Li, and Luo Si. Self-attentive biaffine dependency parsing. In *IJCAI*, pages 5067–5073, 2019. 3, 4

[22] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1841–1851, 2020. 2

[23] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 2

[24] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *CVPR*, pages 10900–10910, 2020. 1

[25] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 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, 2018. 1, 2, 6

[26] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018. 1, 2

[27] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from

text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11592–11601, 2019. 7

[28] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10810–10819, 2020. 1, 2, 6, 7

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3, 6

[30] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 6

[31] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2464–2473, 2020. 1, 2

[32] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45:2673–2681, 1997. 4

[33] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1058, 2016. 1

[34] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, pages 510–526, 2016. 6

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

[36] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015. 1

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 3, 6

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017. 3

[40] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2, 6, 7

[41] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 334–343, 2019. 1, 2

[42] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 284–293, 2019. 5

[43] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019. 2, 6, 7

[44] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic (ACL)*, 2020. 3, 4

[45] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 534–544, 2019. 1, 2, 6, 7

[46] Yitian Yuan, Tao Mei, and Wenwu 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*, volume 33, pages 9159–9166, 2019. 1, 2, 5, 7

[47] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10287–10296, 2020. 1, 2, 6, 7

[48] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1247–1257, 2019. 2

[49] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3, 6, 7

[50] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 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, 2019. 1, 2, 5, 6, 7

[51] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017. 1