Language-Guided Multi-Granularity Context Aggregation for Temporal Sentence Grounding

Guoqiang Gong, Linchao Zhu, Yadong Mu

***Abstract*—Temporal sentence grounding in videos is a crucial task in vision-language learning. Its goal is retrieving a video segment from an untrimmed video that semantically corresponds to a natural language query. A video usually contains multiple semantic events, which are rarely isolated. They tend to be temporally ordered and semantically correlated (*e.g.*, some event is often the precursor of another event). To precisely localize a semantic moment from a video, it is critical to effectively extract and aggregate multi-granularity contextual information, including the fine-grained local context around the moment-related video segment (in short snippet-level) and coarse-grained semantic correlation (in segment-level). Additionally, a second main insight in this work is that the above context aggregation should be favorably guided by the queries, rather than fully query-agnostic. Putting above ideas together, we here present a new network that does language-guided multi-granularity context aggregation. It is comprised of two major modules. The core of the first module is a novel language-guided temporal adaptive convolution (LTAC) devised to extract fine-grained information over video snippets around the ground-truth video segment. It decomposes a convolution into two channel-oriented / temporal-oriented ones. In particular, the convolutional channels are supposed to be more susceptible to queries, thus we learn to generate a dynamic channel-oriented kernel with respect to the querying sentence. As a second module, we propose a language-guided global relation block (LGRB) that extracts video-level context. It augments the contextual feature by using a multi-scale temporal attention that tackles the scale variation of ground-truth video segments, and a multi-modal semantic attention that relies on syntactic of the query. For the validation purpose, we have conducted comprehensive experiments on two popularly-adopted video benchmarks (*i.e.*, ActivityNet Captions and Charades-STA). All experimental results and ablation studies have clearly corroborated the effectiveness of our model designs, outstripping prior state-of-the-art methods in terms of major performance metrics for the task.**

***Index Terms*—Vision-language learning, video understanding, temporal sentence grounding, multi-modality learning.**

I. Introduction

With the tremendous consumer videos aggregated over social networks, surveillance systems and personal albums, intelligent video analysis techniques have attracted increasing attention from both academia and industry. Localizing complex activities in untrimmed videos is a fundamental problem in video understanding. A few of previous works have attempted to find the temporal boundary of a number of predefined action categories (*e.g.*, the sports action "high jump"), referred to as temporal action localization. Recently, increasing research efforts have been devoted to a more challenging setting dubbed as *temporal sentence grounding*, where a natural language sentence is used to describe complex activities more flexibly. For an untrimmed video, temporal sentence grounding aims to localize a most likely segment in the video, which contains the activity that semantically matches the language sentence.

In the literature of temporal sentence grounding, a large body of existing methods have adopted a two-stage strategy [1]–[4]. They first generate video segment-level proposals (*e.g.*, using the sliding window) across an untrimmed video, and then rank a candidate window by its similarity with the query sentence. We would emphasize that most of these methods process different proposals separately and do not fully exploit the global context of videos. One-stage methods have also been widely explored, which directly regress the temporal boundary. Existing works either utilize temporal convolution [5], [6] or self-attention mechanism [7]–[9] for obtaining non-local receptive field. For proposal generation, anchor-based methods [5], [6], [10] pre-specify a set of templates with different temporal locations and scales. In contrast, anchor-free methods [7], [11], [12] directly predict the temporal boundaries of target segments.

A consumer video typically encapsulates a number of complex events that collectively weave the same storyline. One of the main observations in this work is that different events in the same video are often not isolated. The occurrence of some event can provide temporal cues when attempting to localize another event. As seen from an example in Fig. 1, the event "cleaning snow from his car" is a strong hint that an event "opens the doors of his car and gets in" is likely to occur in the following. When a query sentence is issued as the input, exploiting the full landscape of the video semantics, rather than focusing on specific segment-level proposal, is more favored in order to elevate the temporal localization accuracy. In addition, an event often gradually fades into the background, making a temporal localization model suffering from the event-background confusion. To precisely estimate the temporal boundary for a query sentence in an untrimmed video, it is also crucial to collect sufficient fine-grained local context from the video frames temporally around the ground-truth video segment.

Conventionally, features of different modalities in the task of temporal sentence grounding are separately extracted before

Guoqiang Gong and Yadong Mu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: gonggq@pku.edu.cn, myd@pku.edu.cn). Yadong Mu is also affiliated with Peng Cheng Laboratory. Part of the work was performed when the first author was an intern at Baidu Research. The research is supported by Science and Technology Innovation 2030 - New Generation Artificial Intelligence (2020AAA0104401), Beijing Natural Science Foundation (Z190001), and Peng Cheng Laboratory Key Research Project No.PCL2021A07.

Linchao Zhu is with the School of Computer Science, Zhejiang University, Hangzhou, China (e-mail: zhulinchao@zju.edu.cn).

Corresponding author: Yadong Mu.

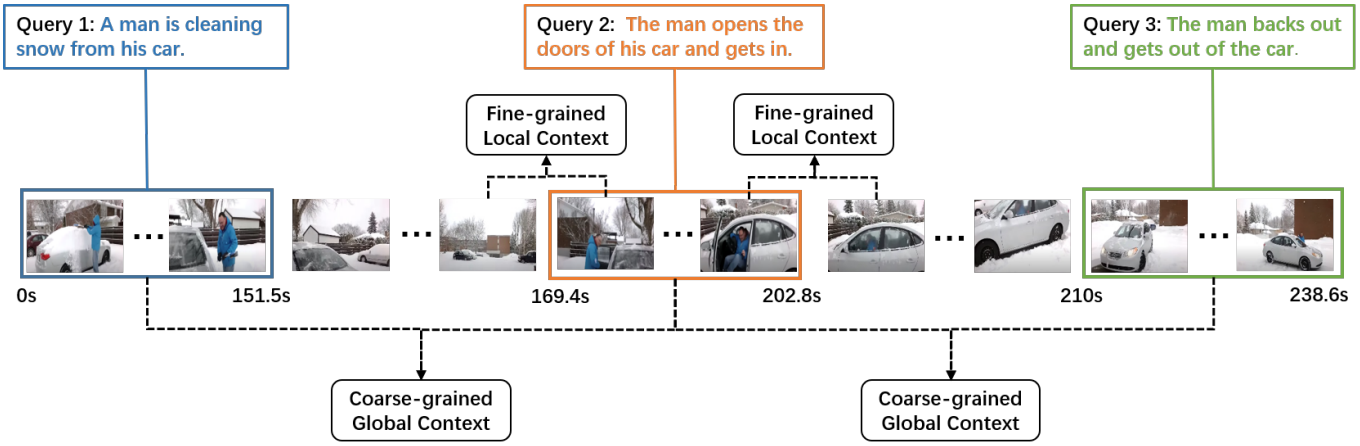

Fig. 1: An illustrating example for the temporal sentence grounding task. Activities in a video have temporal dependencies and semantic relationships between each other. Precisely locating the target video segment corresponding to query sentence (such as Query 2) requires leveraging the query for modeling coarse-grained activity relation and fine-grained local discrimination.

the critical cross-modal matching, as exemplified in the work of 2D-TAN [13]. Inspired by a more recent work SMIN [14], this work advocates language-based guidance for extracting multi-granularity contextual information. Intuitively, the semantics of a video or image are often multi-faceted. For example, images from the Microsoft COCO benchmark can be described by 5 different natural language captions complementary to each other. We postulate a proactive scheme for feature extraction that harnesses the guidance of queries can significantly fill the gap between vision-language modalities.

This paper proposes a language-guided multi-granularity context aggregation network for the temporal sentence grounding task. It dynamically aggregates the snippet-level fine-grained local context and video-level coarse-grained global context guided by the semantic and syntactic information of the natural language query. The fine-grained local context around the ground-truth video segment provides discriminative information for temporal boundary estimation. In contrast, the coarse-grained global context within the entire video captures the relationship between video segments and further promotes the match between the target segment and query sentence. Correspondingly, the proposed network includes two core modules.

- To encode the fine-grained local context, we present a *language-guided temporal adaptive convolution* (LTAC) that is composed of both temporal and channel-oriented convolutional kernels. In specific, The temporal kernel is location-sensitive, which aims to aggregate local context information dynamically. The channel kernel is location-invariant and generates the query-semantic related video representation. Importantly, the weights of temporal and channel kernels are dynamically predicted from the input video and sentence features, respectively.
- To encode the coarse-grained global context, we propose a *language-guided global relation block* (LGRB) which implements multi-scale temporal attention and multi-modal semantic attention. As shown in Fig. 1, video segments corresponding to different queries have notable temporal scale variations. To tackle it, multi-scale temporal attention is used to model the long-term temporal dependencies between activities. The multi-modal semantic attention aims at modeling the semantic relationship between activities with the help of sentence syntactic information.

Finally, the fine-grained local context and coarse-grained global context are fused adaptively to construct the appropriate context vectors for each video frame. To validate the effectiveness of the proposed network, we have conducted comprehensive empirical evaluations on two representative video benchmarks widely adopted in the task of temporal sentence grounding (*i.e.*, ActivityNet Captions and Charades-STA). The experimental results consistently demonstrate large performance gains in almost all settings in comparison with previous state-of-the-art methods.

II. Related Work

A. Temporal Action localization

Temporal action localization [15]–[22] aims to find the temporal boundary of an instance of some pre-defined action from untrimmed videos. The development of temporal action localization techniques has been heavily influenced by the visual object detection methods in images, such as Faster-RCNN. A majority of temporal action localization methods [17], [19] utilize a two-stage pipeline. In practice, they first generate a short-list of action proposals, and afterwards rank these action proposals with adjusted temporal boundaries.

For the proposal-generating stage, existing solutions can be roughly divided into sliding window methods [17], [23]–[26], boundary point detection methods [27]–[31], and temporal actionness grouping [19], [32]. S-CNN [17] generates windows of varying lengths to detect the actions at different temporal scales. R-C3D [26] proposes a fully 3D convolution network to extract video features and predicts the relative offset for predefined anchors to generate proposals. TAL-Net [20] adopts the Faster R-CNN pipeline and proposes a network with different receptive fields and late feature fusion

to improve the quality of proposals. Instead of using a region proposal network like Faster R-CNN, SSN [19] generates proposals by grouping continuous temporal regions with high actionness scores. CTAP [32] proposes a proposal-level actionness trustworthiness estimator to fuse the sliding window and temporal actionness grouping method. BSN [27] detects temporal boundaries with high confidence scores and combines candidate starting and ending boundaries into proposals. BMN [28] further improves the effectiveness of the BSN by introducing an end-to-end Boundary-Matching mechanism. TSA [30] proposes temporal convolution with various dilation rates to ensemble temporal context information at different scales. BUTAL [31] proposes a consistency regularization to improve the accuracy of the boundary and inner points detection. In the second stage, the temporal boundaries of proposals are refined and assigned to action categories. SSN [19] models the temporal structure of each proposal by temporal pyramid. PGCN [33], RAM [34], and G-TAD [35] consider the relations between the proposals and generated a more informative feature representation for each proposal.

In addition to the two-stage methods, there are also one-stage methods [21], [24], [36], [37] that directly localize the action instances. While great progress has been achieved, these methods are limited to locating pre-defined action categories, such as long jump, drinking water, etc. To overcome this limitation, the temporal sentence grounding task is proposed.

B. Temporal Sentence Grounding in Videos

Temporal sentence grounding [1], [2], [38]–[44] aims to locate the temporal boundaries in an untrimmed video of an event described by a querying sentence. Similar to temporal action localization, early temporal sentence grounding methods adopt a two-stage scheme. CTRL [1] and MCN [2] use sliding window methods to generate proposals with different lengths. CTRL [1] utilizes the multi-modal processing module to predict the proposal matching score and adjust the temporal boundaries. MCN [2] projects the proposal feature and sentence feature into a shared embedding space, then uses Euclidean distance to rank proposals. Similar to the region proposal network in object detection, QSPN [45] designed a query-guided segment proposal network to generate query-related proposals.

Despite the effectiveness of two-stage architecture, these methods generate too many overlapped proposals, resulting in high computational overhead. Inspired by the one-stage object detection methods such as SSD [46], there are other threads of methods that directly generate grounding results in a single pass. SCDM [5] utilizes a semantic modulated hierarchical temporal convolution network to generate anchors with different temporal lengths. The overlapping score and location offsets for each anchor are predicted based on modulated features. 2D-TAN [13] and SMIN [14] use a 2D temporal map to represent diverse video moments. SMIN disentangles the activity moment into boundary / content and performs cross-modal interaction coupled with structured moment interaction. However, SMIN neglects the temporal scale variations of activities and ignores the syntactic structure of sentences. Different from SMIN, we propose multi-scale temporal attention to handle activities of different temporal scales and multi-model semantic attention to aggregate multi-modal semantic context by referring to both the syntactic and semantic information of a sentence. MS-2D-TAN [47] further improves the efficiency and performance of the 2D-TAN by proposing a multi-scale 2D temporal map. Some anchor-free methods such as LGI [7] and DRN [11] directly regress the temporal boundaries of video moments without generating anchors. To generate multi-granularity video representation, LGI [7] employs a residue convolution block and a Non-Local block [48] to capture local and global context, respectively. DRN [11] obtains hierarchical feature maps by constructing the feature pyramid. However, both methods generate multi-granularity video representation without considering the semantic and syntactic structure of the query sentence. Such query-agnostic multi-granularity video representation might not be suitable for locating the target video segments corresponding to different query sentences. Different from previous methods, we propose a language-guided temporal adaptive convolution and a language-guided global relation block to generate query-related multi-granularity video representation. Besides, there also exist some works [49]–[52] that adopt the framework of reinforcement learning. These methods formulate the temporal boundary localization of the target segment as a sequential decision-making process.

Unlike recent studies, we consider the multi-granularity context modeling under the guidance of query sentences. We fuse fine-grained local and coarse-grained global contexts to construct the appropriate context vectors for each video snippet flexibly to generate precise grounding results.

III. The Proposed Method

This section will first briefly introduce the problem formulation, followed by detailed description of the proposed model. The overview of our model is illustrated in Fig. 2. As seen, it consists of four major components: (1) language and video encoder. (2) language-guided multi-granularity context aggregation. (3) proposal generation block. (4) localization block. We also elaborate on the training and inference details.

A. Problem Formulation

Given a query sentence S and an untrimmed video V, the goal of temporal sentence grounding is to localize the best video segment that semantically matches the query sentence. More specifically, we denote the query sentence as a sequence of words $S = \{w_n\}_{n=1}^{N}$, where w_n is the n-th word and N is the length of sentence. For an untrimmed video V, we first segment it into T video snippets, with each snippet as a composition of a few consecutive frames. The untrimmed video can then be denoted as a sequence of snippets $V = \{v_t\}_{t=1}^{T}$, where v_t is the t-th snippet.

B. Language and Video Encoding

For a query sentence S, an embedding vector is generated for each word using GloVe [53]. Then, the word embedding vectors are fed as input into a two-layer bi-directional

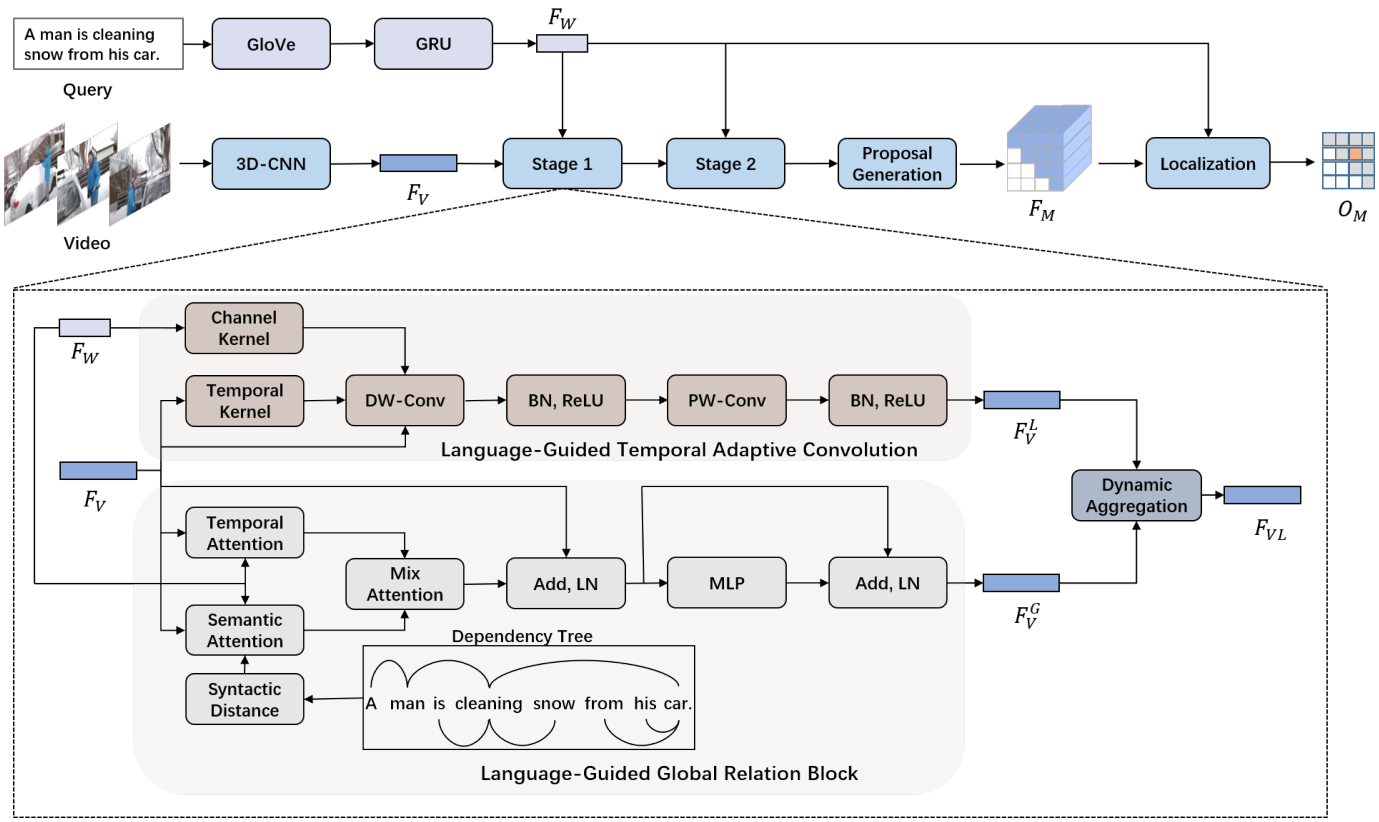

Fig. 2: The overall architecture of our proposed model. We first extract the word-level query's features and video features with GRU and 3D-CNN, respectively. Then the coarse-grained global context and fine-grained local context are calculated in two successive language-guided multi-granularity context aggregation stages (Stage 1 and Stage 2) and generate the multi-granularity context-aware feature. Both stages have the same architecture, consisting of a language-guided temporal adaptive convolution (LTAC), a language-guided global relation block (LGRB), and a dynamic aggregation block. Afterwards, we generate proposal features based on the context-aware feature in proposal generation block. Finally, a localization block predicts the video segment semantically corresponding to the query.

GRU [54] to obtain contextual feature representations. Let $F_W = \{f_{w_n}\}_{n=1}^{N} \in \mathbb{R}^{N \times D_q}$ be the concatenation of hidden states of GRU along both directions. The sentence feature $F_S \in \mathbb{R}^{1 \times D_q}$ is the average of F_W.

Following the practice in previous works [7], [14], [47], a pre-trained CNN model is used to extract features for each video snippet. The features of all snippets are concatenated along the temporal dimension to get the video feature sequence $F_V = \{f_{v_t}\}_{t=1}^{T} \in \mathbb{R}^{T \times D_v}$ where D_v denotes the dimension of features. Furthermore, we apply two linear layers to the language and video features respectively, reducing the dimensions to D_c.

C. Language-Guided Temporal Adaptive Convolution

It frequently occurs that the predicted segment only partially overlaps with the ground-truth video segment to a query sentence, rendering a low temporal IoU (intersection-over-union) score. To precisely estimate the temporal boundary, it desires to investigate fine-grained temporal context information, particularly the local context extracted from video snippets around the true boundary. Previous works [7], [11] adopt standard temporal convolution to extract local context. The weights of standard temporal convolution are shared across all the input video/query pairs. Given the varied nature of contents in a video and query sentence, a shared temporal convolution kernel seems sub-optimal to capture temporal context for the temporal sentence grounding task. Therefore, we predict the dynamic kernel weights from the input video and sentence that provides custom parameters for different video/query pairs.

To this end, the proposed model incorporates language-guided temporal adaptive convolution (LTAC). For efficacy consideration, LTAC is designed to decouple the temporal / feature channels. State differently, it is composed of a temporal kernel and a channel kernel. The former is dependent on the time stamp of a video snippet, assigning different weights according to the temporal location in the video. Instead, the latter kernel is designed to be invariant for all time stamps in the video. Since there will be some noises in the video boundary context information that is irrelevant to the query sentence. To mitigate the negative effects of noises, we generate the channel kernel F_C conditioned on sentence feature to extract the query-semantic related local context.

Recall that both the video feature F_V and sentence feature F_W are transformed into D_c-dimensional. Let k denotes

the size of temporal and temporal and channel kernels. To generate the channel kernel, we first transform the sentence feature $F_W \in \mathbb{R}^{N\times D_c}$ into feature map $F_{CK} \in \mathbb{R}^{N\times D_c}$ and $F_{CW} \in \mathbb{R}^{N\times k}$ by two independent transformations, respectively. More specifically, F_W is transformed into F_{CK} by $F_{CK} = F_W W_K$, where $W_K \in \mathbb{R}^{D_c\times D_c}$ is a learnable matrix. Each channel of F_{CK} is supposed to partially capture the semantics of the query sentence. F_W is transformed into F_{CW} via $F_{CW} = \text{Softmax}(F_W W_C)$, where $W_C \in \mathbb{R}^{D_c\times k}$ is a learnable matrix, Softmax is performed along the first dimension to normalize the weights. F_{CW} contains k groups of attention weights over words of a sentence. Each group of attention weights can capture a different semantic aspect of the query sentence. The channel kernel F_C is generated by estimating k groups of attention weights over words of a sentence and computing a weighted sum of the transformed sentence feature as below:

$$F_C = F_{CW}^{\intercal} F_{CK} \in \mathbb{R}^{k\times D_c}. \tag{1}$$

Equation 1 is related to the factorized bilinear pooling methods [55]–[58]. Different from factorized bilinear pooling, we first transform the sentence feature into attention weights F_{CW} and feature map F_{CK}, then generate channel kernel via attention-weighted sum operation.

To aggregate local context for each snippet dynamically, we generate the temporal kernel by

$$F_T = \text{Softmax}\left(\text{Conv1D}(F_V^{\intercal})\right) \in \mathbb{R}^{k\times T}, \tag{2}$$

where Softmax is performed along the first dimension to normalize the temporal kernel weights. the temporal and channel kernels are combined to perform a depth-wise convolution (DW-Conv) as below:

$$F_{VD}[t,c] = \sum_{o=-r}^{r} F_T[o+r,t]\cdot F_C[o+r,c]\cdot F_V[o+t,c], \tag{3}$$

where $r = floor(\frac{k}{2})$. c denotes the feature channel index for the t-th snippet. A point-wise convolution (PW-Conv) then applies a 1-D convolution with kernel size 1 to create a linear combination of F_{VD}, and generates the semantic-aware local context feature F_V^L. Batch normalization (BN) and ReLU are appended after each convolutional layer.

D. Language-Guided Global Relation Block

Next, to fully exploit the temporal ordering and semantic correlation among different events in a video, we further propose a language-guided global relation block (LGRB). The core of LGRB include a multi-scale temporal attention (MSTA) branch and a multi-modal semantic attention (MMSA) branch.

1) Multi-Scale Temporal Attention: Video events typically exhibit large variation in terms of duration lengths. MSTA is the major workhorse in our model for tackling the temporal scale issue. We first generate multi-scale video features by applying several temporal pooling operations with different output sizes in parallel. Since not all video frames are relevant to the query sentence, a sentence guided pooling (SGP) operation will filter out the irrelevant video information during the pooling process. In practice, F_S is first converted into $\hat{F}_S \in \mathbb{R}^{1\times D_c}$ by a fully connected layer. Then, we perform Sigmoid over the inner-product between $\hat{F}_S$ and F_V, obtaining the query sentence-guided pooling weight P_W as below:

$$P_W = \text{Sigmoid}\left(\frac{F_V \hat{F}_S^{\intercal}}{\sqrt{D_c}}\right) \in \mathbb{R}^{T\times 1}. \tag{4}$$

Let f_{ts}^{te} be the output of SGP performed on a video segment which ranges from ts-th snippet to te-th snippet, SGP is defined as:

$$f_{ts}^{te} = \frac{\sum_{t=ts}^{te} P_W[t]F_V[t]}{\sum_{t=ts}^{te} P_W[t]} \in \mathbb{R}^{1\times D_c}. \tag{5}$$

When $P_W[t] = 1, t = 1\ldots T$, SGP boils down to the average pooling. Let M be the number of pooling operations. A video is divided into T_m segments during the m-th ($1 \le m \le M$) pooling process. $P_{V,m} = \text{SGP}_m(F_V) \in \mathbb{R}^{T_m\times D_c}$ is the output of the m-th pooling operation. The multi-scale feature P_V is calculated by concatenating the outputs of all pooling operations along the temporal dimension:

$$P_V = \text{Concat}\left(P_{V,1}, P_{V,2}, \ldots, P_{V,M}\right) \in \mathbb{R}^{(\sum_{m=1}^{M} T_m)\times D_c}. \tag{6}$$

In the experiments, we use 3 pooling operations and set T_1, T_2, T_3 as 1, 8, 16, respectively. Motivated by the multi-head self-attention [59], multi-scale temporal attention consists of H attention heads. For each head h, we project F_V and P_V into $Q_{Th} = F_V W_{QTh}$, $K_{Th} = P_V W_{KTh}$, and $V_{Th} = P_V W_{VTh}$, where $W_{QTh}, W_{KTh}, W_{KTh} \in \mathbb{R}^{D_c\times D_h}$ represent the weights of projection layer. $D_h = D_c/H$ is the feature dimension of each head. Then, Q_{Th}, K_{Th} and V_{Th} are used to calculate the multi-scale temporal attention F_{VTh} of head h according to:

$$F_{VTh} = \text{Softmax}\left(\frac{Q_{Th}\times K_{Th}^{\intercal}}{\sqrt{D_h}}\right)\times V_{Th}. \tag{7}$$

The output of different attention heads are concatenated along the channel dimension to get multi-scale temporal attention $F_{VT} \in \mathbb{R}^{T\times D_c}$ as:

$$F_{VT} = \text{Concat}(F_{VT1}, F_{VT2}, ..., F_{VTH}). \tag{8}$$

Compared with the original multi-head self-attention, multi-scale temporal attention is more efficient since the temporal length of P_V is less than F_V. In addition, multi-scale temporal attention can model relationships between activities with different duration since P_V contains multi-scale information.

2) Multi-modal Semantic Attention: In order to model the multi-modal semantic context, we first calculate the feature similarity $C \in \mathbb{R}^{T\times N}$ between the query words and snippets as:

$$C = \frac{(F_V W_{V1})(F_W W_{W1})^{\intercal}}{\sqrt{D_c}}, \tag{9}$$

where W_{V1} and W_{W1} are learnable parameter matrices. $C[t,n]$ denotes the similarity between the t-th snippet and n-th word. Although some words do not appear in the video, their contextual feature representations still contain video-related semantic information due to our adoption of bi-directional GRU during language encoding. To generate the attention on

each snippet with respect to each word, we normalize C over the first dimension to obtain a word-specific snippet attention matrix $\hat{C} = \text{Softmax}(C)$. We aggregate the video features with respect to each word based on $\hat{C}$ by:

$$F_{VW} = \hat{C}^\intercal F_V \in \mathbb{R}^{N \times D_c}, \tag{10}$$

where $F_{VW}[n]$ represents the visual representation related to the n-th word. The visual and language representations are combined to get multi-modal $F_{MW} \in \mathbb{R}^{N \times D_c}$ of N words as:

$$F_{MW} = F_{VW} W_{VW} + F_W W_{W2}, \tag{11}$$

where W_{VW} and W_{W2} are learnable matrices with their sizes inferred from the context.

While $F_{MW}[n]$ contains the multi-modal information of the n-th word, it lacks the semantic context from relevant words. To aggregate multi-modal semantic context, we propagate information among multi-modal word representations with the help of dependency parsing tree of the query sentence. Specifically, we first obtain the syntactic dependency tree G for a sentence by Stanford CoreNLP toolkit. In G, each word is regarded as a node and each dependency relation between a word pair is a directed edge. To enable bi-directional message passing between words, we convert edges to be un-directed. Intuitively, the distance between two words in the dependency tree can reflect their linguistic affinity. If words w_i and w_j are connected by at least a path, the syntactic distance $d_{i,j}$ between w_i and w_j is set to the length of the shortest path. Otherwise the syntactic distance between two words is infinity. Let $E \in \mathbb{R}^{N \times N}$ be the affinity matrix of N words. $E[i,j]$ is calculated by combining the feature similarity and syntactic distance as

$$E[i,j] = \frac{F_{MW}[i] F_{MW}[j]^\intercal}{d_{i,j}}. \tag{12}$$

For numerical tractability, E is further normalized such that the sum of each row is 1. We calculate the semantic context aware multi-modal word representation P_M as follows:

$$P_M = (E + I) F_{MW} W_{MW} \in \mathbb{R}^{N \times D_c}, \tag{13}$$

where I is an identity matrix serving as a shortcut connection to ease the optimization. $W_{MW} \in \mathbb{R}^{D_c \times D_c}$ is a learned matrix. Similar to MSTA, MMSA also adopts a multi-head design. For head h, F_V and P_M are projected into $Q_{Sh} = F_V W_{QSh}$, $K_{Sh} = P_M W_{KSh}$, and $V_{Sh} = P_M W_{VSh}$, where $W_{QSh}, W_{KSh}, W_{KSh} \in \mathbb{R}^{D_c \times D_h}$ represent the weights of the projection layer. Then MMSA of head h is calculated as:

$$F_{VSh} = \text{Softmax}\left(\frac{Q_{Sh} \times K_{Sh}^\intercal}{\sqrt{D_h}}\right) \times V_{Sh}, \tag{14}$$

where $F_{VS} \in \mathbb{R}^{T \times D_c}$ is the concatenation of all heads' outputs.

3) Mixed Attention: To integrate information from MSTA and MMSA, we first concatenate F_{VT} and F_{VS} along the channel dimension, then generate the mixed attention F_{VM} by $F_{VM} = \text{Concat}(F_{VT}, F_{VS}) W_{VM} \in \mathbb{R}^{T \times D_c}$, where $W_{VM} \in \mathbb{R}^{(D_c + D_c) \times D_c}$ is a learned matrix. Afterwards, the coarse-grained global context F_V^G is calculated by:

$$\hat{F}_{VM} = \text{LN}(F_{VM} + F_V), \tag{15}$$

$$F_V^G = \text{LN}(\text{MLP}(\hat{F}_{VM}) + \hat{F}_{VM}), \tag{16}$$

where LN denotes the layer normalization and MLP consists of two linear transformations.

E. Dynamic Aggregation Block

The importance of local and global context may vary across different video snippets. We are thus motivated to propose a snippet-adaptive fusion module to aggregate the local and global context. Specifically, we first transform F_V^L and F_V^G into F_V^{L1} and F_V^{G1} by two independent Conv1D layers with kernel size 1, respectively. We further concat the F_V^{L1} and F_V^{G1} along the channel dimension and use a two-layer MLP to produce the selection weights SW for local and global context feature as:

$$SW = \text{Softmax}(\text{MLP}(\text{Concat}(F_V^{L1}, F_V^{G1}))) \in \mathbb{R}^{T \times 2}, \tag{17}$$

where Softmax is performed along the last dimension of SW to get the normalized selection weights. The multi-granularity context-aware feature F_{VL} is obtained as follows:

$$F_{VL} = SW[:,0] \odot F_V^{L1} + SW[:,1] \odot F_V^{G1}. \tag{18}$$

F. Proposal Generation Block

For a video containing T snippets, enumerating all possible video segments is computationally expensive. To reduce the computation cost, we first down-sample F_{VL} to $\hat{F}_{VL} \in \mathbb{R}^{T_D \times D_c}$ where $T_D \ll T$ by temporal average pooling. Then, we construct 2D proposal feature map [13] F_M with a boundary-matching operation [28]:

$$F_M = \text{BM}(\hat{F}_{VL}) \in \mathbb{R}^{T_D \times T_D \times D_c}. \tag{19}$$

For a valid proposal (a, b) which satisfies $0 \le a \le b < T_D$, its starting and ending moments are $\frac{a}{T_D} V_d$ and $\frac{b+1}{T_D} V_d$ respectively, with V_d being the duration of the video.

G. Localization Block

After obtaining the proposal feature map, we then predict the matching score of each valid proposal. Inspired by 2D-TAN [13], we interact the sentence feature F_S with the proposal feature map F_M through Hadamard product and $\mathcal{L}_2$ norm as:

$$F_{MF} = \mathcal{L}_2 Norm(F_M W_M \odot F_S W_{S1}), \tag{20}$$

where $W_M, W_{S1} \in \mathbb{R}^{D_c \times D_c}$ are learnable parameter matrices to project F_M and F_S to the same subspace.

For the fused proposal feature map F_{MF}, each row has the same starting time and each column has the same ending time. To aggregate the context information from other proposals, we first perform average pooling along each row and each column separately to get $F_{MF}^r \in \mathbb{R}^{T_D \times D_c}$ and $F_{MF}^c \in \mathbb{R}^{T_D \times D_c}$. After that, F_{MF}^r and F_{MF}^c are fed into two Conv1D layers with kernel size 3 to aggregate proposal-level

local context. Then, we integrate F^r_{MF} and F^c_{MF} to generate the proposal context feature map $\hat{F}_{MF} \in \mathbb{R}^{T_D \times T_D \times D_c}$, where $\hat{F}_{MF}[a,b,:] = F^r_{MF}[a,:] + F^c_{MF}[b,:]$. The context-augmented proposal feature map F_{MC} is generated by:

$$F_{MC} = F_{MF} + \text{Sigmoid}(\text{Conv1D}(\hat{F}_{MF}))F_{MF}. \quad (21)$$

The 2D proposal matching score map O_M is predicted by:

$$O_M = \text{Sigmoid}(F_{MC}W_{OM}) \in \mathbb{R}^{T_D \times T_D \times 1}, \quad (22)$$

where $W_{OM} \in \mathbb{R}^{D_c \times 1}$ denotes a learnable parameter matrix. O_M is flattened into a matching score sequence. A valid matching score stand for $O_M[a,b]$, where $0 \leq a \leq b < T_D$. All valid matching scores are collected, denoted as $\hat{O}_M \in \mathbb{R}^C$, where C is the number of valid proposals.

Besides the matching score map, we also predict the temporal boundary of an event and the actionness of each snippet to facilitate the network training. The actionness O_A of each snippet is predicted as:

$$O_A = \text{Sigmoid}\left(\frac{F_{VL}(F_S W_{S2})^\intercal}{\sqrt{D_c}}\right) \in \mathbb{R}^{T \times 1}, \quad (23)$$

where $W_{S2} \in \mathbb{R}^{D_c \times D_c}$ is a learnable parameter matrix. The label of the actionness of the i-th snippet is set to 1 when the i-th snippet is in the ground truth video segment and 0 otherwise. The probability sequence of starting boundary O_S is generated by:

$$O_S = \text{Softmax}[(\text{ReLU}(F_{VL}W_{OS1}))W_{OS2}] \in \mathbb{R}^{T \times 1}, \quad (24)$$

where $W_{OS1} \in \mathbb{R}^{D_c \times D_c/2}$, and $W_{OS2} \in \mathbb{R}^{D_c/2 \times 1}$ are learnable parameter matrices. The probability sequence of ending boundary O_E is likewise computed.

H. Training and Inference

Following previous practice [13], we use a scaled IoU value as the supervision signal of the predicted matching score. Specifically, for the i-th proposal, we first calculate its IoU o_i with the ground truth video segment. Then the scaled IoU value y_i is assigned by

$$y_i = \begin{cases} 0 & o_i \leq t_{min}, \\ \frac{o_i - t_{min}}{t_{max} - t_{min}} & t_{min} < o_i < t_{max}, \\ 1 & o_i \geq t_{max}, \end{cases} \quad (25)$$

where t_{min} and t_{max} are two thresholds. The proposal matching loss is defined as

$$\mathcal{L}_m = -\frac{1}{C}\sum_{i=1}^{C} y_i \log \hat{O}_M[i] + (1 - y_i)\log(1 - \log \hat{O}_M[i]). \quad (26)$$

To supervise the predicted actionness of each video snippet, we use the temporal actionness loss as:

$$\mathcal{L}_{ta} = -\frac{1}{T}\sum_{i=1}^{T} G_A[i]\log(O_A[i]) + (1 - G_A[i])(1 - \log(O_A[i])), \quad (27)$$

where $G_A[i]$ is set to 1 if the i-th snippet is in the ground truth video segment and 0 otherwise. In addition, the boundary prediction loss is defined as the Kullback-Leibler divergence between the predicted and ground truth boundary probability distributions as below:

$$\mathcal{L}_b = \text{KL}(O_S || G_S) + \text{KL}(O_E || G_E), \quad (28)$$

where G_S and G_E are the starting and ending boundary probability distributions, respectively.

The training of our model is supervised by:

$$\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_{ta} + \beta\mathcal{L}_b, \quad (29)$$

where α and β are the weighting coefficients.

During inference, we rank all moment proposals based on the proposal matching score. Non-maximum suppression (NMS) is used to remove duplicate proposals.

TABLE I: Comparisons on ActivityNet Captions using C3D features. The best / second best scores under each setting are highlighted in red and green respectively. "-" implies that the corresponding score was not reported in the original paper.

Method	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
MCN [2]	21,26	6.43	53.23	29.70
TGN [60]	28.47	-	43.33	-
CTRL [1]	29.01	10.34	59.17	37.54
ACRN [3]	31.67	11.25	60.34	38.57
QSPN [45]	33.26	13.43	62.39	40.78
CBP [61]	35.76	17.80	65.89	46.20
SCDM [5]	36.75	19.86	64.99	41.53
ABLR [62]	36.79	-	-	-
GDP [9]	39.27	-	-	-
RWM-RL [49]	36.90	-	-	-
TSP-PRL [50]	38.76	-	-	-
LGI [7]	41.51	23.07	-	-
2D-TAN [13]	44.51	26.54	77.13	61.96
CMIN [10]	43.40	23.88	67.95	50.73
DRN [11]	45.45	24.36	77.97	50.30
CBLN [63]	48.12	27.60	79.32	63.41
MSATN [64]	48.02	31.78	78.02	63.18
CPN [65]	45.10	28.10	-	-
IVG [66]	43.84	27.10	-	-
DeNet [67]	43.79	-	74.13	-
FVMR [68]	45.00	26.85	77.42	61.04
SSCS [69]	46.67	27.56	78.37	63.78
MS-2D-TAN [47]	46.93	28.23	78.79	61.29
Ours	48.88	32.88	79.43	65.80

IV. Evaluations

A. Dataset and Evaluation Metric

a) ActivityNet Captions [70]: ActivityNet Captions is originally constructed for dense video captioning. It consists of about 20K untrimmed videos and is divided into training, validation and testing sets at a ratio of 2:1:1. The average duration of videos is around 120 seconds. Each video contains 3.65 querying sentences and corresponding video moments on average. Following the common practice as in [14], [63], [64], we use *val_1* as the validation set and *val_2* as the testing set since annotations of the testing set are not publicly available.

b) Charades-STA [1]: Charades-STA re-purposes the Charades dataset for promoting the video grounding research. Most of the videos in Charades depict human daily indoor activities. The average length of the videos is about 30 seconds. This dataset consists of 12,408 and 3,720 moment-query pairs in the training and testing set, respectively.

c) Evaluation Metric: As in previous work [1], [14], "R@n, IoU=m" is used as the major evaluation metric. The metric denotes the percentage of testing samples that have at least one correct grounding prediction (*i.e.*, the IoU between the prediction and the ground truth is larger than m) in the top-n predictions. In the experiments, we set $n \in \{1, 5\}$ and $m \in \{0.5, 0.7\}$ for ActivityNet Captions and Charades-STA.

B. Implementation Details

For fair comparison, we use C3D [71] network pre-trained on Sports-1M [72] to extract video features on ActivityNet Captions and Charades-STA. Following previous work [5], [63], we also use I3D [73] network pre-trained on Kinetics [73] to extract video features on the Charades-STA. To encode each word in a sentence, we use Glove word embedding with 300 dimensions. A two-layer bidirectional GRU is applied to word-embeddings to obtain the word and sentence feature representation. During training, We use Adam with a learning rate of 0.0001, the momentum of 0.9 and batch size of 32. For the scaled IoU, the scaling thresholds are set to 0.5 and 1.0 for ActivityNet Captions and Charades-STA, respectively. The weighting coefficients α and β are set to 1 and 0.1 for both datasets.

C. Comparisons with State-of-the-art Methods

Table I presents the results on ActivityNet Captions. Our method outperforms other state-of-the-art temporal sentence grounding methods in terms of most metrics, demonstrating the superiority of our method. Compared with MSATN [64], our method outperforms it with a margin of 2.6% under R@5,IoU=0.7. Since IoU=0.7 is a more rigorous criterion to determine whether a localized moment is correct or not, this demonstrates that the proposed method generates grounding results with superior quality.

Table II summarizes the evaluations on Charades-STA. As seen, our method achieves the best performance under all IoU thresholds. We observe that the improvement is more significant at high IoU. Specifically, for R@1, IoU=0.7, our method elevates the performance from 40.75% to 43.70%, using I3D features. When using C3D feature, our method outperforms MS-2D-TAN [47] by 3% and 6.54% under R@1,IoU=0.7 and R@5,IoU=0.7, respectively.

The comparisons in Tables I and II are comprehensive since various types of proposal-generating methods are involved. For sliding window based methods, we choose the representative CTRL [1], MCN [2], ACRN [3], and ACL-K [4] for comparison. The inferior performances of these methods are supposed to be partly owing to treating each proposal independently and ignoring the temporal context. There are also anchor based methods including SCDM [5], TGN [60], and CMIN [10]. Anchor-based methods conduct sentence language interactions and learn temporal context by temporal convolution or recurrent neural network. However, they only exploit snippet-level context, while our method exploits the fine-grained local context and coarse-grained global context simultaneously to generate grounding results with more precise temporal boundaries. We also compare anchor free methods such as DRN [11], LGI [7] and DeNet [67]. Due to the target video segments having various temporal duration and semantics, directly regressing the temporal boundary is difficult. Therefore, the grounding results of anchor-free methods are usually not accurate enough.

TABLE II: Comparisons with state-of-the-art methods on Charades-STA. The best / second best scores under each setting are highlighted in red and green respectively. "-" implies that the corresponding score was not reported in the original paper.

Method	Feature	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CTRL [1]	C3D	23.63	8.89	58.92	29.57
ACL-K [4]	C3D	30.48	12.20	64.84	35.13
RWM-RL [49]	C3D	36.70	-	-	-
TSP-PRL [50]	C3D	37.39	17.69	-	-
QSPN [45]	C3D	35.60	15.80	79.40	45.40
CBP [61]	C3D	36.80	18.87	70.94	50.19
GDP [9]	C3D	39.47	18.49	-	-
FVMR [68]	C3D	38.16	18.22	82.18	44.96
MS-2D-TAN [47]	C3D	41.10	23.25	81.53	48.55
Ours	C3D	44.10	26.26	84.36	55.09
DRN [11]	I3D	53.09	31.75	89.06	60.05
SCDM [5]	I3D	54.44	33.43	74.43	58.08
LGI [7]	I3D	59.46	35.48	-	-
CBLN [63]	I3D	61.13	38.22	90.33	61.69
CPN [65]	I3D	59.77	36.67	-	-
SMIN [14]	I3D	64.06	40.75	89.49	68.09
DeNet [67]	I3D	59.70	38.52	91.24	66.83
FVMR [68]	I3D	55.01	33.74	89.17	57.24
LGI + SSCS [69]	I3D	60.75	36.19	-	-
MS-2D-TAN [47]	I3D	60.08	37.39	89.06	59.17
Ours	I3D	64.13	43.70	91.30	69.15

TABLE III: Ablation studies of main components.

Dataset	Method	R@1, IoU=0.5	R@1 IoU=0.7
Charades-STA	w/o LGRB	60.23	39.78
	w/o LTAC	61.45	41.90
	w/o LGRB & LTAC	51.21	32.70
	Full	64.13	43.70
ActivityNet	w/o LGRB	46.84	30.26
	w/o LTAC	46.79	31.11
	w/o LGRB & LTAC	44.39	27.05
	Full	48.88	32.88

D. Ablation Studies

a) Effectiveness of main components: We evaluate the main components of the proposed method on Charades-STA and ActivityNet Captions in Table III, where "w/o LGRB" means without language-guided global relation block (LGRB),

TABLE IV: Ablation studies of components of language-guided temporal adaptive convolution (LTAC).

Dataset	Method	R@1, IoU=0.5	R@1 IoU=0.7
Charades-STA	Conv1D	51.32	33.02
	LTAC w/o CK	52.82	34.15
	LTAC w/o TK	58.06	38.47
	LTAC	60.23	39.78
ActivityNet	Conv1D	44.91	27.10
	LTAC w/o CK	45.44	28.01
	LTAC w/o TK	46.09	28.93
	LTAC	46.84	30.26

TABLE V: Ablation studies of components of language-guided global relation block (LGRB).

Dataset	Method	R@1, IoU=0.5	R@1 IoU=0.7
Charades-STA	w/o MSTA	59.74	40.63
	w/o MMSA	55.58	36.74
	LGRB	61.45	41.90
ActivityNet	w/o MSTA	46.29	30.31
	w/o MMSA	45.97	29.12
	LGRB	46.79	31.11

"w/o LTAC" means without language-guided temporal adaptive convolution (LTAC), and "w/o LGRB & LTAC" means without both LGRB and LTAC. LGRB and LTAC are two critical components in our model, which conduct coarse-grained global context for semantic reasoning and fine-grained local context for precise boundary localization. We observe that the performance of our method degenerates dramatically without LGRB or LTAC. The full model (Full) outperforms all the compared ablation models. These facts demonstrate that LGRB and LTAC are complementary for temporal sentence grounding.

b) Variants of language-guided temporal adaptive convolution (LTAC): We validate the design of the LTAC in Table IV by comparing four settings: 1)"Conv1D" means standard temporal convolution 2) "LTAC w/o CK" means without channel kernel. 3) "LTAC w/o TK" means without temporal kernel. 4) "LTAC" means use full LTAC in our model. From Table IV, we observe that both temporal kernel and channel kernel contribute to the overall performance. The channel kernel is critical by comparing "w/o CK" with "LTAC", which indicates the sentence semantics play an important role in this task. LTAC consistently outperforms Conv1D on two datasets, demonstrating the effectiveness of the dynamic kernel weights used in LTAC.

c) Variants of language-guided global relation block (LGRB): We verify the effectiveness of multi-scale temporal attention (MSTA) and multi-modal semantic attention (MMSA) in Table V by comparing three settings: 1) "w/o MSTA" means LGRB without MSTA. 2) "w/o MMSA" means LGRB without MMSA. 2) "LGRB" means the full LGRB with both MSTA and MMSA. By comparing "LGRB" with "w/o MSTA" and "w/o MMSA", we observe that both multi-scale temporal context and multi-modal semantic context contribute to the overall performance. From Table V, removing MMSA from LGRB degrades R@1,IoU=0.7 significantly on both Charades-STA(41.90% vs. 36.74%) and ActivityNet Captions (31.11% vs 29.12%).

TABLE VI: Ablation studies of the kernel size k of LTAC on Charades-STA.

Kernel Size	R@1, IoU=0.5	R@1 IoU=0.7
1	57.75	38.62
3	60.23	39.78
5	59.42	39.59
7	59.53	39.26

TABLE VII: Impact of the number of pooling operations M and the length of each pooling operation of SGP on the Charades-STA dataset.

M	$\{T_m, 1 \leq m \leq M\}$	R@1, IoU=0.5	R@1 IoU=0.7
1	8	58.40	39.18
	16	58.72	39.88
	32	59.04	40.26
3	1, 2, 4	59.90	41.09
	1, 4, 8	60.74	41.68
	1, 8, 16	61.45	41.90
	1, 16, 32	60.92	41.66
5	1, 2, 4, 8, 16	61.01	41.81

d) Kernel size k in LTAC: Table VI shows the impact of k on Charades-STA. We observe that compared with setting kernel size as 1, increasing kernel size to 3 improves the R@1, IOU=0.5 from 57.75% to 60.23%, which shows the importance of fine-grained temporal context information. We further note that when the kernel size is larger than 3, the performance deteriorates a little. In our experiments, the k is set to 3.

e) Parameter choices of SGP: To validate the effectiveness of using multiple temporal pooling operations and figure out the optimal hyper-parameters of SGP, we use different combinations of M and $\{T_m, 1 \leq m \leq M\}$. The results are shown in Table VII. As can be seen, our model achieves better performance when using multiple pooling operations. This demonstrates that multi-scale temporal attention is helpful for temporal sentence grounding. We further note that using 3 pooling operations is enough to achieve good results, and adding more pooling operations does not improve performance. SGP with $M = 3$ and $T_1 = 1, T_2 = 8, T_3 = 16$ achieves the highest performance as shown in Table VII, which is applied in our experiments.

f) Different temporal pooling methods of multi-scale temporal attention: To evaluate the temporal pooling methods in the multi-scale temporal attention, we conduct ablation studies on Charades-STA concerning R@1 in Fig. 3. "None" means without pooling, same to multi-head self-attention [59], "Avg" means average pooling. "Max" means max pooling. "SGP" means semantic guided pooling. The results show that SGP performs better than other pooling methods since the

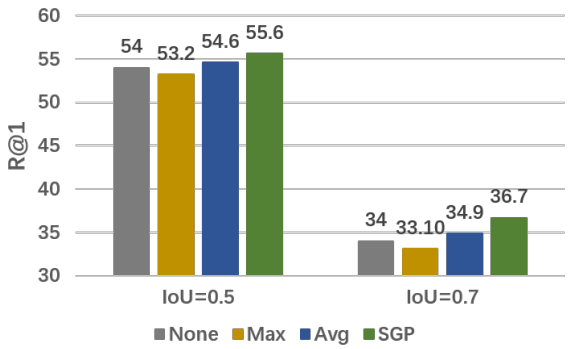

Fig. 3: Ablation studies of different pooling methods in multi-scale temporal attention.

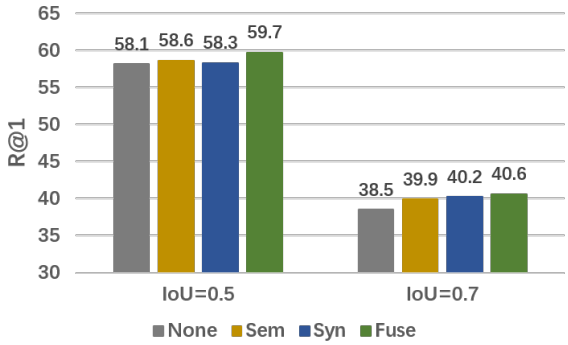

Fig. 4: Ablation studies of different affinity matrices in multi-modal semantic attention.

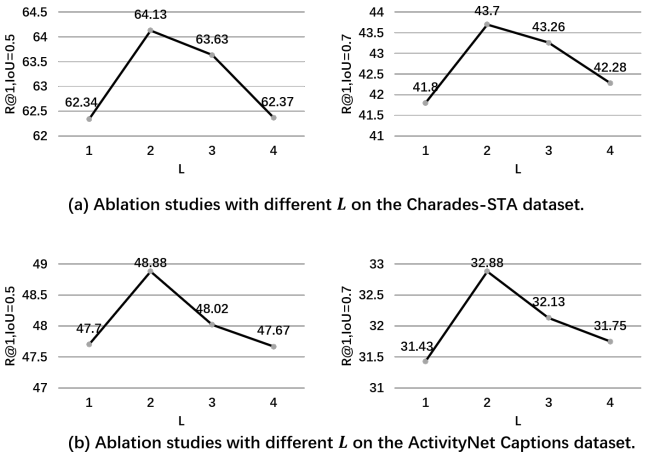

Fig. 5: Ablation studies of the number of stages L in our model.

TABLE VIII: Ablation studies of different loss.

Dataset	Method	R@1, IoU=0.5	R@1 IoU=0.7
Charades-STA	w/o $\mathcal{L}_{ta}$	61.65	41.87
	w/o $\mathcal{L}_{b}$	60.21	41.28
	w/o $\mathcal{L}_{ta}$&$\mathcal{L}_{b}$	58.67	39.56
	Full	64.13	43.70
ActivityNet	w/o $\mathcal{L}_{ta}$	46.67	31.02
	w/o $\mathcal{L}_{b}$	46.99	31.09
	w/o $\mathcal{L}_{ta}$&$\mathcal{L}_{b}$	45.78	30.25
	Full	48.88	32.88

semantic related video information is preserved during the pooling process. We also notice that the performance of "Max" is even worse than "None". This may be due to the loss of semantic related information during the max pooling process.

g) Different affinity matrix in multi-modal semantic attention: To evaluate the affinity matrix in multi-modal semantic attention, we conduct ablation studies on Charades-STA in Fig. 4. "None", "Sem", "Syn", "Fuse" denote that the affinity matrix is an identity matrix, calculated by the feature similarity / sentence syntactic distance, or combining the feature similarity and sentence syntactic distance, respectively. It is seen that the semantic context is beneficial for this task by comparing "Fuse" to "None". As presented in Fig. 4, the result of "Fuse" is better than "Sem" and "Syn", indicating that the feature similarity and sentence syntactic distance are complementary to each other.

h) Number of Multi-Granularity Context Aggregation Stage: We show the effect of the count of stages L in Fig. 5. L varies from 1 to 4 in the proposed model. The results show that our model achieves the best performance when using two stages. We observe that increasing the stage L from 2 to 4 does not improve the performance. Adding more layers will increase the over-fitting risk.

i) Effectiveness of each loss function: We also evaluate the loss functions used in the proposed method on Charades-STA and ActivityNet Captions in Table VIII, where "w/o $\mathcal{L}_{ta}$" means our model is trained without temporal actionness loss. "w/o $\mathcal{L}_{tb}$" means our model is trained without boundary prediction loss. "w/o $\mathcal{L}_{ta}$&$\mathcal{L}_{b}$" means our model is trained by proposal matching loss. The model trained by full loss (Full) outperforms all the compared ablation models. This demonstrates that the temporal actionness loss and boundary prediction loss can facilitate the model learning, and the three losses are complementary for temporal sentence grounding.

E. Qualitative Analysis

We show some qualitative examples in Fig. 6, where "w/o LGRB" and "w/o LTAC" correspond to models without language-guided global relation block or language-guided temporal adaptive convolution respectively, respectively. As seen, the full model predicts more precise boundaries than ablation models, since the full model captures both the coarse-grained global context and fine-grained local context.

V. Conclusions

In this paper, we study the temporal sentence grounding task and present a novel language-guided multi-granularity context aggregation network. We design a novel language-guided temporal adaptive convolution to extract fine-grained information over video snippets and a language-guided global relation block to extract video-level context. Our method incorporates both fine-grained local discriminative information

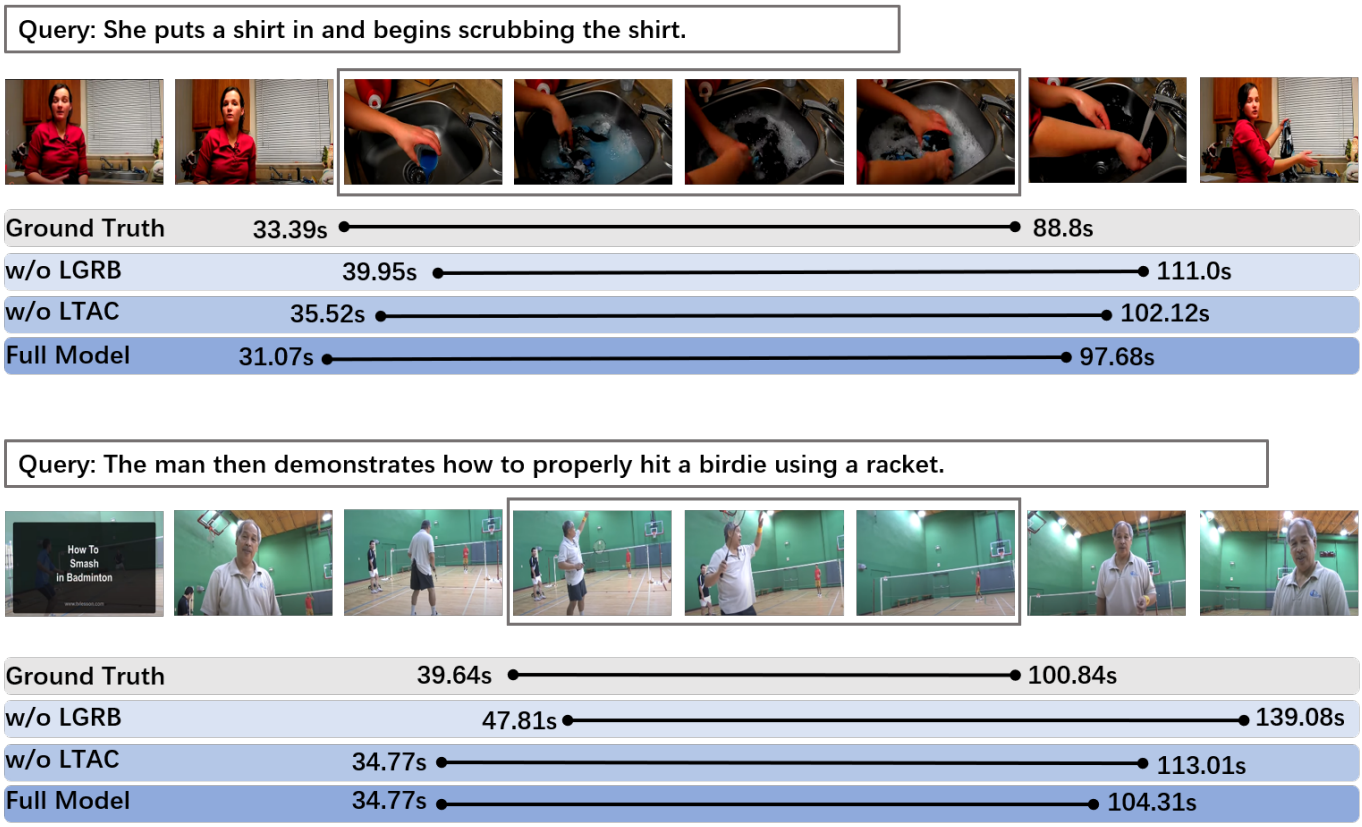

Fig. 6: Qualitative examples of our model and the ablation models.

and coarse-grained global semantic relation to locate the target video segment precisely. Comprehensive experimental results on the ActivityNet Captions and Charades-STA datasets show that our work achieves the new state-of-the-art performance.

REFERENCES

[1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5267–5275. 1, 3, 7, 8

[2] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell, "Localizing moments in video with natural language," in *IEEE International Conference on Computer Vision*, 2017, pp. 5804–5813. 1, 3, 7, 8

[3] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018, pp. 15–24. 1, 7, 8

[4] R. Ge, J. Gao, K. Chen, and R. Nevatia, "MAC: mining activity concepts for language-based temporal localization," in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 245–253. 1, 8

[5] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *Advances in Neural Information Processing Systems*, 2019, pp. 534–544. 1, 3, 7, 8

[6] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257. 1

[7] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 810–10 819. 1, 3, 4, 7, 8

[8] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6543–6554. 1

[9] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, and X. Li, "Rethinking the bottom-up framework for query-based video localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 10 551–10 558. 1, 7, 8

[10] Z. Zhang, Z. Lin, Z. Zhao, and Z. 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*, 2019, pp. 655–664. 1, 7, 8

[11] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 287–10 296. 1, 3, 4, 7, 8

[12] C. Lu, L. Chen, C. Tan, X. Li, and J. Xiao, "DEBUG: A dense bottom-up grounding approach for natural language video localization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 5143–5152. 1

[13] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 870–12 877. 2, 3, 6, 7

[14] H. Wang, Z.-J. Zha, L. Li, D. Liu, and J. Luo, "Structured multi-level interaction network for video moment localization via language query," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7026–7035. 2, 3, 4, 7, 8

[15] G. Gong, X. Wang, Y. Mu, and Q. Tian, "Learning temporal co-attention models for unsupervised video action localization," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 9816–9825. 2

[16] G. Gong, L. Zheng, W. Jiang, and Y. Mu, "Self-supervised video action localization with adversarial temporal transforms," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Z. Zhou, Ed., 2021, pp. 693–699. 2

[17] Z. Shou, D. Wang, and S. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1049–1058. 2

[18] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "Cdc:

Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1417–1426. 2

[19] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2933–2942. 2, 3

[20] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139. 2

[21] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 344–353. 2, 3

[22] J. Li, X. Liu, Z. Zong, W. Zhao, M. Zhang, and J. Song, "Graph attention based proposal 3d convnets for action detection," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 4626–4633. 2

[23] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 768–784. 2

[24] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: single-stream temporal action proposals," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6373–6382. 2, 3

[25] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia, "TURN TAP: temporal unit regression network for temporal action proposals," in *IEEE International Conference on Computer Vision*, 2017, pp. 3648–3656. 2

[26] H. Xu, A. Das, and K. Saenko, "R-C3D: region convolutional 3d network for temporal activity detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 5794–5803. 2

[27] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: boundary sensitive network for temporal action proposal generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–21. 2, 3

[28] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: boundary-matching network for temporal action proposal generation," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3888–3897. 2, 3, 6

[29] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S. Chang, "Multi-granularity generator for temporal action proposal," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3604–3613. 2

[30] G. Gong, L. Zheng, and Y. Mu, "Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos," in *IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6. 2, 3

[31] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian, "Bottom-up temporal action localization with mutual regularization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 539–555. 2, 3

[32] J. Gao, K. Chen, and R. Nevatia, "CTAP: complementary temporal action proposal generation," in *Computer Vision - ECCV 2018 - 15th European Conference*, 2018, pp. 70–85. 2, 3

[33] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7093–7102. 3

[34] P. Chen, C. Gan, G. Shen, W. Huang, R. Zeng, and M. Tan, "Relation attention for temporal action localization," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2723–2733, 2020. 3

[35] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, "G-tad: Sub-graph localization for temporal action detection," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 153–10 162. 3

[36] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 988–996. 3

[37] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *British Machine Vision Conference*, 2017. 3

[38] S. Zhang, J. Su, and J. Luo, "Exploiting temporal relationships in video moment localization with natural language," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1230–1238. 3

[39] B. Liu, S. Yeung, E. Chou, D.-A. Huang, L. Fei-Fei, and J. Carlos Niebles, "Temporal modular networks for retrieving complex compositional activities in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 569–586. 3

[40] H. Wang, Z.-J. Zha, X. Chen, Z. Xiong, and J. Luo, "Dual path interaction network for video moment localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4116–4124. 3

[41] C. R. Opazo, E. Marrese-Taylor, F. S. Saleh, H. Li, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2464–2473. 3

[42] D. Liu, X. Qu, X.-Y. Liu, J. Dong, P. Zhou, and Z. Xu, "Jointly cross-and self-modal graph attention network for query-based moment localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4070–4078. 3

[43] D. Cao, Y. Zeng, X. Wei, L. Nie, R. Hong, and Z. Qin, "Adversarial video moment retrieval by jointly modeling ranking and localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 898–906. 3

[44] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, and J. Xiao, "Boundary proposal network for two-stage natural language video localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2986–2994. 3

[45] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9062–9069. 3, 7, 8

[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 21–37. 3

[47] S. Zhang, H. Peng, J. Fu, Y. Lu, and J. Luo, "Multi-scale 2d temporal adjacency networks for moment localization with natural language," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3, 4, 7, 8

[48] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803. 3

[49] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8393–8400. 3, 7, 8

[50] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 12 386–12 393. 3, 7, 8

[51] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 334–343. 3

[52] D. Cao, Y. Zeng, M. Liu, X. He, M. Wang, and Z. Qin, "Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4162–4170. 3

[53] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543. 3

[54] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734. 4

[55] Y. Li, N. Wang, J. Liu, and X. Hou, "Factorized bilinear models for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2098–2106. 5

[56] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1839–1848. 5

[57] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, 2018. 5

[58] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," in *International Conference on Learning Representations*, 2017. 5

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. 5, 9

[60] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, 2018, pp. 162–171. 7, 8

[61] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 168–12 175. 7, 8

[62] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9159–9166. 7

[63] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng, W. Wei, Z. Xu, and Y. Xie, "Context-aware biaffine localizing network for temporal sentence grounding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 235–11 244. 7, 8

[64] M. Zhang, Y. Yang, X. Chen, Y. Ji, X. Xu, J. Li, and H. T. Shen, "Multi-stage aggregated transformer network for temporal language localization in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 669–12 678. 7, 8

[65] Y. Zhao, Z. Zhao, Z. Zhang, and Z. Lin, "Cascaded prediction network via segment tree for temporal video grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4197–4206. 7, 8

[66] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2765–2775. 7

[67] H. Zhou, C. Zhang, Y. Luo, Y. Chen, and C. Hu, "Embracing uncertainty: Decoupling and de-bias for robust temporal grounding," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2021, pp. 8445–8454. 7, 8

[68] J. Gao and C. Xu, "Fast video moment retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1523–1532. 7, 8

[69] X. Ding, N. Wang, S. Zhang, D. Cheng, X. Li, Z. Huang, M. Tang, and X. Gao, "Support-set based cross-supervision for video grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 573–11 582. 7, 8

[70] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-Captioning Events in Videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 706–715. 7

[71] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497. 8

[72] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732. 8

[73] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308. 8

Guoqiang Gong received his B.E. degree from Sun Yat-Sen University, Guangzhou, China, in 2018. He is currently a Ph.D. candidate at the Wangxuan Institute of Computer Technology, Peking University. His research interests include video understanding and temporal action localization.

Linchao Zhu (Member, IEEE) received the B.E. degree from Zhejiang University, China, in 2015, and the Ph.D. degree in computer science from the University of Technology Sydney, Australia, in 2019. He is a Research Professor with the College of Computer Science and Technology, Zhejiang University, China. His research interests are video analysis and understanding.

Yadong Mu is an Assistant Professor at Wangxuan Institute of Computer Technology, Peking University. He obtained both the B.S. and Ph.D. degrees from Peking University. Before joining Peking University, he had ever worked as research fellow at National University of Singapore, research scientist at Columbia University, researcher at Huawei Noah's Ark Lab in Hong Kong, and senior scientist at AT&T Labs. His research interest is in broad research topics in computer vision and machine learning.