Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos

Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, *Senior Member, IEEE* and Wenwu Zhu, *Fellow, IEEE*

Abstract—Temporal sentence grounding in videos aims to localize one target video segment, which semantically corresponds to a given sentence. Unlike previous methods mainly focusing on matching semantics between the sentence and different video segments, in this paper, we propose a novel semantic conditioned dynamic modulation (SCDM) mechanism, which leverages the sentence semantics to modulate the temporal convolution operations for better correlating and composing the sentence-relevant video contents over time. The proposed SCDM also performs dynamically with respect to the diverse video contents so as to establish a precise semantic alignment between sentence and video. By coupling the proposed SCDM with a hierarchical temporal convolutional architecture, video segments with various temporal scales are composed and localized. Besides, more fine-grained clip-level actionness scores are also predicted with the SCDM-coupled temporal convolution on the bottom layer of the overall architecture, which are further used to adjust the temporal boundaries of the localized segments and thereby lead to more accurate grounding results. Experimental results on benchmark datasets demonstrate that the proposed model can improve the temporal grounding accuracy consistently, and further investigation experiments also illustrate the advantages of SCDM on stabilizing the model training and associating relevant video contents for temporal sentence grounding.

Index Terms—Temporal sentence grounding in videos (TSG), semantic conditioned dynamic modulation (SCDM), temporal convolution.

1 Introduction

Detecting or localizing activities in videos [1], [2], [3], [4], [5], [6], [7], [8], [9] is a prominent while fundamental problem for video understanding. As videos often contain intricate activities that cannot be indicated by a predefined list of action classes, a new task, namely temporal sentence grounding in videos (TSG) [10], [11], has recently attracted much research attention [12], [13], [14], [15], [16], [17], [18]. Formally, given an untrimmed video and a natural sentence query, the TSG task aims to identify the start and end timestamps of one specific video segment, which contains activities of interest semantically corresponding to the given sentence query.

Most of existing approaches [10], [11], [15], [16] for the TSG task often sample candidate video segments first, then fuse the sentence and video segment representations together, and thereafter evaluate their matching relationships based on the fused features. Lately, some approaches [12], [13] try to directly fuse the sentence information with each video clip, then employ an LSTM or a ConvNet to compose the fused features over time, and finally predict the temporal boundaries of the target video segment with boundary regressors. While promising results have been achieved, there are still several problems that need to be considered.

First, previous methods mainly focus on semantically matching sentences and individual video segments or clips, while neglect the important guiding role of sentences to help correlate and compose video contents over time. For example, the target video sequence shown in Fig. 1 mainly expresses two distinct activities "*woman walks cross the room*" and "*woman reads the book on the sofa*". Without referring to the sentence, these two distinct activities are not easy to be associated as one whole event. However, the sentence clearly indicates that "*The woman takes the book across the room to read it on the sofa*". Keeping such a semantic meaning in mind, persons can easily correlate the two activities together and thereby precisely determine the temporal boundaries. Therefore, how to make use of the sentence semantics to guide the composing and correlating of relevant video contents over time is crucial for the TSG task.

Second, video contents are usually of various visual appearances over time. Hence, the sentence guidance for video content composition and segment boundary determination should also dynamically evolve with the diverse visual appearances, so as to detect target segments in videos in a more flexible and sensitive way.

Third, the video activities that are specified by query sentence often have unconstrained temporal locations and scales. However, when matching or composing video segments with sentence queries over time, the initially obtained segments of the previous methods are regularly distributed and manually defined over the video sequence [11], [15]. Even though the temporal boundaries of these segments are adjusted by the temporal boundary regression network in previous work [11], [15], the adjustment is still limited in the regions around the predefined segments, yielding temporal grounding results with imprecise boundaries. Therefore, it is worthy of consideration that how to effectively use the sentence information to adjust and refine the temporal boundaries of the localized segment.

To solve the above problems for the TSG task, in this paper, we propose a novel semantic conditioned dynamic

- *Y. Yuan is with the Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China.*
- *W. Zhu is with the Department of Computer Science and Technology, Tsinghua University, China.*
- *L. Ma, J. Wang, and W. Liu are with Tencent AI Lab, China.*

Manuscript received XXXX; revised XXXX.

Fig. 1. The temporal sentence grounding in videos (TSG) task. Our proposed SCDM relies on the sentence to modulate the temporal convolution operations, which can thereby establish the sentence-video semantic interaction in a light-weight manner, and temporally correlate and compose the various sentence-relevant activities (highlighted in red and green) to get the grounding results. Additionally, the SCDM-coupled temporal convolutional network can also predict the actionness scores over the video sequence, and help adjust the boundaries of the segment for more accurate grounding results.

modulation (SCDM) mechanism, which leverages sentence semantic information to modulate the temporal convolution process in a hierarchical temporal convolutional architecture. The SCDM manipulates the temporal feature maps by adjusting the scaling and shifting parameters for feature normalization with referring to the sentence semantics. As such, the temporal convolution process is modulated and activated to better associate and compose sentence-relevant video contents over time. By coupling the proposed SCDM with a hierarchical temporal convolutional architecture, our model naturally characterizes the interaction behaviors between sentence and video segments of various temporal scales, and further predicts multi-scale segment-level proposals for grounding the sentence query. Moreover, in order to provide more precise temporal boundaries for the grounding proposals, one novel clip-level actionness prediction module, which predicts the sentence specified actionness scores over time in a more fine-grained clip-level [7], [19], can evaluate whether a clip is the starting/ending/middle point of the target localized segment.

We conduct experiments on three benchmark datasets, and the experimental results show the superiority of the proposed SCDM-coupled temporal convolutional architecture over the state-of-the-art methods. The clip-level actionness prediction module also demonstrates its benefits of further improving the boundary accuracy of the predicted temporal video segments. Additionally, we also conduct experiments to investigate the model performances when SCDM is imposed on different temporal convolutional layers, the effects of SCDM for model learning, and the influence of SCDM on correlating temporal video features.

To summarize, our main contributions are as follows. (1) We propose a novel semantic conditioned dynamic modulation (SCDM) mechanism, which dynamically modulates the temporal convolution procedure by referring to the sentence semantic information. By doing so, the sentence-relevant video contents can be temporally correlated and composed to yield a precise temporal boundary prediction. (2) Coupling the proposed SCDM with a hierarchical temporal convolutional network, our model naturally exploits the complicated semantic interactions between sentence and video in various temporal granularities in a lightweight manner. The SCDM-based segment-level proposal prediction along with the clip-level actionness prediction further makes the temporal grounding procedure comprehensive and precise. (3) Experiments conducted on three public datasets verify the effectiveness of the proposed SCDM mechanism as well as our overall architecture. Moreover, the advantage of SCDM in stabilizing the model training, its good convergence, and its ability of guiding the composition of sentence-relevant video contents are further demonstrated.

2 Related Works

2.1 Temporal Sentence Grounding in Videos

Early works for temporal sentence grounding in videos mainly constrain to certain visual scenarios (videos in the laboratory, movie or kitchen environment), or aim at aligning multiple sentences within a single video [20], [21], [22], [23]. Recently, Gao [11] and Hendricks *et al.* [10] generalized the task into more general cases. Hendricks *et al.* [10] took the multi-modal matching strategy to solve the temporal grounding problem. They embedded video segment and sentence query features into a common latent space, where the distances between matched segment and query pairs are minimized. Gao *et al.* [11] and Liu *et al.* [15] adapted the object detection methodologies from spatial domains to temporal domains, and established a location regression network based on the multi-modal fusion features of video segments and sentence queries, hence predicting the temporal offsets between candidate video segments and ground-truth segments. Ge *et al.* [24] further extended the location regression method by mining activity concepts from both videos and queries to enhance the temporal grounding task. One common issue of the above methods is that they all need to sample several candidate segments of various temporal scales from videos via sliding window searching, which will inevitably cause redundant computation.

Going beyond the sliding window sampling, Xu *et al.* [25] proposed to generate query-specific proposals from videos first, and then performed triplet ranking among these proposals by learning both a fine-grained similarity metric for retrieval and query re-generation. Chen *et al.* [16] also integrated the semantic information of natural sentence queries into a proposal generation process, and then yielded visual concept vectors by a learned visual concept detection CNN. Video frame features and query sentence features were jointly utilized to evaluate and refine the boundaries of the previously generated proposals. Overall, both of these two methods separate the proposal generation procedure with the proposal ranking or refinement procedure, making the overall grounding system barely fully optimized.

There are another thread of works that get rid of explicit candidate segment generation, and directly generate temporal grounding results in a single pass. Chen and Ma

et al. [12] dynamically matched sentence and video units via a sequential LSTM grounder, and at each time step, the grounder would score a group of candidate segments with different temporal scales ending at this time. Zhang *et al.* [13] adopted the Graph Convolutional Network (GCN) [26] to model sentence-specified relations among candidate segments produced from a convolutional neural network, where the candidate video segments/moments with different temporal locations and scales are further be ranked in a single shot and their moment-wise temporal relations can be learned jointly. Rodriguez *et al.* presented a proposal-free approach for temporal grounding [27], in which a dynamic filter is leveraged to transfer language information to the visual domain. A new loss function was introduced to guide the model to attend the most relevant parts of the video, and soft labels applied to individual video clips were used to model annotation uncertainty. All of these methods only model the multi-modal interaction between natural sentence queries and fine-grained video clips, while the temporal prediction is performed at various scales of video segments. Therefore, the complex matching relationships between queries and video contents of different granularities or scales are overlooked, which will influence the grounding accuracy of the models.

Although promising results have been achieved by existing methods, they all focus on better aligning semantic information between sentence and video, while neglect the fact that sentence information plays an important role in correlating the described activities in videos. Our work firstly introduces the sentence information as a critical prior to compose and correlate video contents over time. Subsequently, sentence-guided video composing is dynamically performed and evolved in a hierarchical temporal convolution architecture, in order to accommodate the diverse video contents of various temporal granularities.

2.2 Weight Generating Networks in Convolutional Architectures

Our proposed SCDM leverages the sentence information to adjust the scaling and shifting parameters of feature normalization in the convolutional networks. There are also some related works [28], [29], [30], [31], [32], which dynamically generate weights for convolutional architectures conditioned on the model inputs, so as to increase the model capacity or adaptability.

Xu *et al.* introduced the the Dynamic Filter networks (DF) [28], where the filters are generated dynamically conditioned on the inputs. A wide variety of filtering operations can be learned this way, including local spatial transformations, but also others like selective (de)blurring or adaptive feature extraction. For visual question answering, Noh *et al.* proposed a dynamic parameter layer which output is used as parameters of a fully connected layer [29], and the question representation is taken as the input to the dynamic parameter layer. Yang *et al.* proposed conditionally parameterized convolutions (CondConv) [32], which learns specialized convolutional kernels for each example. Differing from the dynamic filter networks, CondConv does not directly generate the convolutional kernel from the inputs, and instead it keeps a group of kernels and the scalar weights of these kernels are computed conditioned on the inputs. The final convolutional kernel is then determined dynamically by weighted averaging these kernels. Since the kernel is computed only once, then convolved across the input, CondConv can increase the model capacity while still maintain its efficiency in inference.

3 The Proposed Model

In this paper, we propose one novel model to handle the TSG task, as illustrated in Fig. 2. Specifically, as shown in Fig. 2(a), the proposed model contains both segment-level proposal prediction and clip-level actionness prediction. The overall model is realized by a hierarchical temporal convolutional architecture equipped with a novel semantic conditioned dynamic modulation (SCDM) mechanism, which is demonstrated in detail in Fig. 2(b). From bottom to up, segment-level proposals with different temporal lengths are produced in different temporal convolutional layers. More fine-grained clip-level actionness scores are also predicted based on the bottom layer of the temporal convolution architecture, evaluating whether each clip is the starting/ending/middle point of the target localized segment. With such clip-level actionness scores, the segment-level proposal boundaries can be temporally adjusted to generate the final temporal grounding results.

In the following, we will first introduce the modules in the proposed model in detail, namely the multimodal fusion, semantic modulated temporal convolution, segment-level proposal prediction, and clip-level actionness prediction. Then, we will present the training objective function and the temporal boundary adjustment procedure in the inference phase.

3.1 Multimodal Fusion

Given an untrimmed video V and a sentence query S, the TSG task aims to determine the start and end timestamps of one video segment, which semantically corresponds to the given sentence query. In order to perform the temporal grounding, the video is first represented as a feature sequence $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_t, \cdots, \mathbf{v}_T]$, where $\mathbf{v}_t$ represents the feature of the t-th clip in the video. Accordingly, the query sentence is represented as $\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_n, \cdots, \mathbf{s}_N]$, where $\mathbf{s}_n$ denotes the word embedding of the n-th word in the sentence.

For solving the TSG task, one essential problem is to understand both the sentence and video contents, and establish their semantic correlation. As such, after obtaining the video and sentence representations, we concatenate each video clip feature with the average sentence features and then obtain the fused multimodal feature by one fully-connected layer with ReLU activation function:

$$\mathbf{f}_t = \text{ReLU}\left(\mathbf{W}^f\left(\mathbf{v}_t \| \bar{\mathbf{s}}\right) + \mathbf{b}^f\right). \tag{1}$$

Here $\mathbf{W}^f$ and $\mathbf{b}^f$ are the learnable parameters. $\bar{\mathbf{s}}$ denotes the global sentence representation, which can be obtained by simply averaging the word-level sentence representation $\mathbf{S}$, and "$\|$" denotes the concatenation process. With such a multimodal fusion strategy, the yielded representation $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^T \in \mathcal{R}^{T \times d_f}$ captures the interactions between sentence and video clips in a fine-grained manner. The

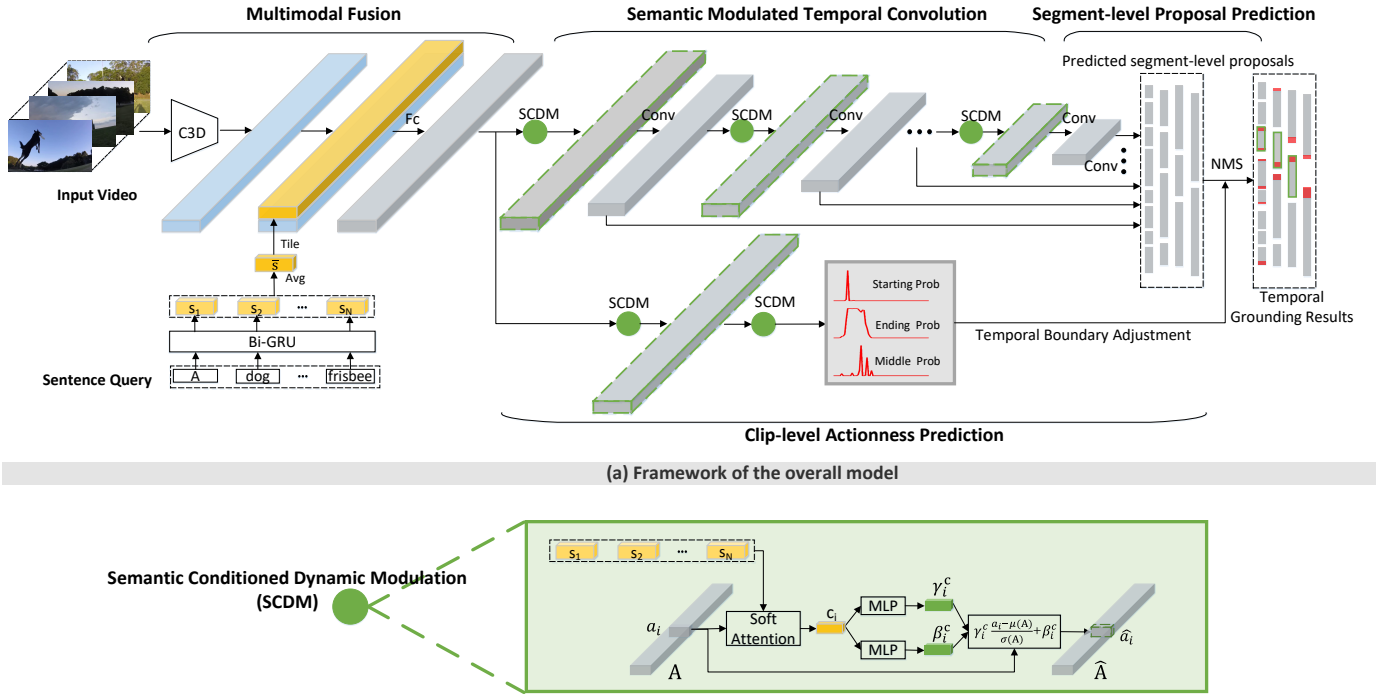


Fig. 2. An overview of our proposed model for the TSG task. Specifically, as shown in the subfigure (a), the multimodal fusion fuses the entire sentence and each video clip in a fine-grained manner. Based on the fused representation, the semantic modulated temporal convolution correlates sentence-relevant video contents in the temporal convolution procedure, with the proposed SCDM dynamically modulating temporal feature maps with respect to the sentence. Afterwards, multiple segment-level proposals are predicted based on the modulated feature maps of different temporal layers. There is another branch of temporal convolution which is imposed on the bottom layer of the convolutional architecture. The clip-level actionness scores are predicted by this branch in a fine-grained level, which are further leveraged to adjust the temporal boundaries of the segment-level proposals, and get the final temporal grounding results. The detailed demonstration of the proposed SCDM mechanism is provided in the subfigure (b). Best viewed in color.

following semantic modulated temporal convolution will gradually correlate and compose such representations over time, expecting to help produce accurate temporal boundary predictions at various scales.

3.2 Semantic Modulated Temporal Convolution

As aforementioned, the sentence-described activities in videos may have various durations and scales. Therefore, the fused multimodal representation $\mathbf{F}$ should be computed from different temporal scales to comprehensively characterize the temporal diversity of video activities. Inspired by the efficient single-shot object and action detections [1], [33], the temporal convolutional network established via one hierarchical architecture is used to produce multi-scale features to cover the activities of various durations. Moreover, in order to fully exploit the guiding role of the sentence, we propose a novel semantic conditioned dynamic modulation (SCDM) mechanism, which relies on the sentence semantics to modulate the temporal convolution operations for better correlating and composing the sentence-relevant video contents over time. In the following, we first review the basics of the temporal convolutional network. Afterwards, the proposed SCDM will be described in detail.

3.2.1 Temporal Convolutional Network

In our proposed model, a typical temporal convolution operation is denoted as $\text{Conv}(\theta_k, \theta_s, d_h)$, where θ_k, θ_s, and d_h indicate the kernel size, stride size, and filter numbers, respectively. Meanwhile, the nonlinear activation, such as ReLU, follows the convolution operation to construct a basic temporal convolutional layer. With the setting $\text{Conv}(3, 1, d_h)$, the output temporal feature map will still keep the same temporal dimension as the input, but each feature unit within the output map will absorb its local context information. With the setting $\text{Conv}(3, 2, d_h)$, the output temporal feature map will halve the temporal dimension of the input feature map and meanwhile expand the receptive field of each feature unit within the map.

Taking the multimodal fusion representation $\mathbf{F}$ as input, we first impose a $\text{Conv}(3, 1, d_h)$ temporal convolutional layer to make an integration of the input representation. Then, by stacking multiple temporal convolutional layers with the setting $\text{Conv}(3, 2, d_h)$, a hierarchical temporal convolutional network is constructed, with each feature unit in one specific feature map corresponding to one specific video segment in the original video. For brevity, we denote the output feature map of the k-th temporal convolutional layer as $\mathbf{A}_k = \{\mathbf{a}_{k,i}\}_{i=1}^{T_k} \in R^{T_k \times d_h}$, $(k = 1, 2, \cdots)$. Here $T_1 = T, T_k = T_{k-1}/2$ $(k = 2, 3, \cdots)$ is the temporal dimension, and $\mathbf{a}_{k,i} \in R^{d_h}$ denotes the i-th feature unit at the the k-th layer feature map. The detailed settings of the temporal convolution architecture will be provided in Sec. 4.2.

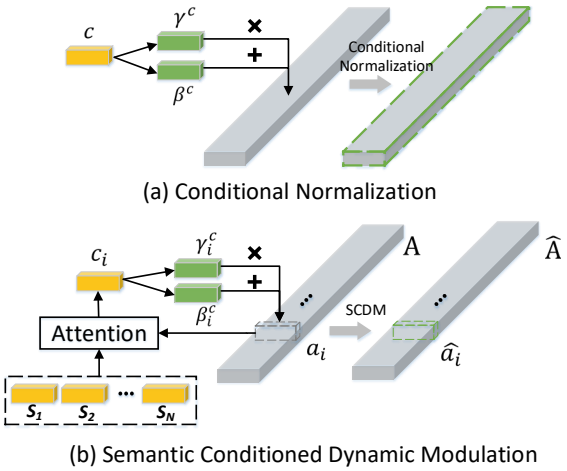


Fig. 3. The comparison between conditional normalization and our proposed semantic conditioned dynamic modulation.

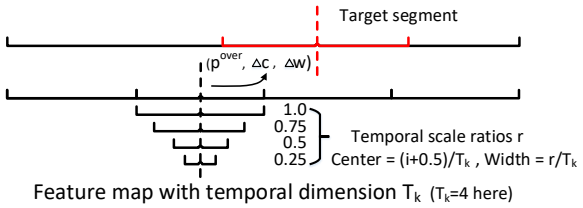


Fig. 4. The illustration of temporal scale ratios and offsets.

3.2.2 *Semantic Conditioned Dynamic Modulation*

Besides the video clip contents, their temporal correlations play an even more important role in video understanding. For the TSG task, as we show in Fig. 1, the query sentence can clearly indicate the temporal correlation among the target activities, and thus provides crucial information to temporally associate and compose the consecutive video contents over time. Based on the above considerations, in this paper, we propose a novel SCDM mechanism, which relies on the sentence semantic information to dynamically modulate the feature composition process in each temporal convolutional layer.

Specifically, as shown in Fig. 3(b), given the sentence representation $\mathbf{S} = \{\mathbf{s}_n\}_{n=1}^N$ and one feature map extracted from one specific temporal convolutional layer $\mathbf{A} = \{\mathbf{a}_i\}$ (we omit the layer number here), we attentively summarize the sentence representation to $\mathbf{c}_i$ with respect to each feature unit $\mathbf{a}_i$:

$$\begin{aligned} \rho_i^n &= \text{softmax}\big(\mathbf{w}^\top \tanh\left(\mathbf{W}^s \mathbf{s}_n + \mathbf{W}^a \mathbf{a}_i + \mathbf{b}\right)\big), \\ \mathbf{c}_i &= \sum_{n=1}^{N} \rho_i^n \mathbf{s}_n, \end{aligned} \tag{2}$$

where $\mathbf{w}$, $\mathbf{W}^s$, $\mathbf{W}^a$, and $\mathbf{b}$ are the learnable parameters. Afterwards, two fully-connected (FC) layers with the `tanh` activation function are used to generate two modulation vectors $\gamma_i^c \in R^{d_h}$ and $\beta_i^c \in R^{d_h}$ that will be used to scale and shift the normalized temporal feature units, respectively:

$$\begin{aligned} \gamma_i^c &= \tanh(\mathbf{W}^\gamma \mathbf{c}_i + \mathbf{b}^\gamma), \\ \beta_i^c &= \tanh(\mathbf{W}^\beta \mathbf{c}_i + \mathbf{b}^\beta), \end{aligned} \tag{3}$$

where $\mathbf{W}^\gamma$, $\mathbf{b}^\gamma$, $\mathbf{W}^\beta$, and $\mathbf{b}^\beta$ are the learnable parameters. Finally, based on the generated modulation vectors γ_i^c and β_i^c, the feature unit $\mathbf{a}_i$ is modulated as:

$$\hat{\mathbf{a}}_i = \gamma_i^c \cdot \frac{\mathbf{a}_i - \mu(\mathbf{A})}{\sigma(\mathbf{A})} + \beta_i^c. \tag{4}$$

With the proposed SCDM, the temporal feature maps, yielded during the temporal convolution process, are meticulously modulated by scaling and shifting the corresponding normalized features under the sentence guidance. This procedure is also analogous to the typical batch/layer normalizations [34] which adjust the normalized feature in the feature space to accelerate the model optimization. In our proposed semantic conditioned dynamic modulation mechanism, each temporal feature map will absorb the sentence semantic information, and further activate the following temporal convolutional layer to better correlate and compose the sentence-relevant video contents over time. Coupling the proposed SCDM with each temporal convolutional layer, we thus obtain the novel semantic modulated temporal convolution as shown in Fig. 2.

Discussion. As shown in Fig. 3, our proposed SCDM differs from the existing conditional batch/instance normalization [35], [36], where the same γ^c and β^c are applied within the whole batch/instance. On the contrary, as indicated in Equations (2)-(4), our SCDM dynamically aggregates the meaningful words with referring to different video contents, making the yielded γ^c and β^c dynamically evolve for different temporal units within each specific feature map. Such a dynamic modulation enables each temporal feature unit to interact with each word to collect useful grounding cues along the temporal dimension. Therefore, the sentence-video semantics can be better aligned over time to support more precise boundary predictions. Detailed experimental demonstrations will be given in Sec. 4.5.

3.3 Segment-level Proposal Prediction

In the hierarchical temporal convolutional architecture, the temporal receptive fields are gradually enlarged while stacking more temporal convolutional layers. Therefore, the lower temporal convolutional layers can be better leveraged to find shorter activities, while the higher layers can be used to localize longer activities. As illustrated in Fig. 4, regarding a feature map with temporal dimension T_k, the basic temporal span for each feature unit within this feature map is $1/T_k$. We impose different scale ratios based on the basic span, and denote them as $r \in R = \{0.25, 0, 5, 0.75, 1.0\}$. As such, for the i-th feature unit of the feature map, we can compute the length of the scaled spans within it as r/T_k, and the center of these spans is $(i + 0.5)/T_k$. For the whole feature map, there are a total number of $T_k \cdot |R|$ scaled spans within it, with each span corresponding to an anchor segment for grounding.

Then, we impose an additional set of convolution operations on the layer-wise temporal feature maps to predict the position of the target video segment. Specifically, each anchor segment will be associated with a prediction vector $p = (p^{over}, \triangle c, \triangle w)$, where p^{over} is the predicted overlap score between the anchor and ground-truth segment, and

$\triangle c$ and $\triangle w$ are the temporal center and width offsets of the anchor segment relative to the ground-truth. Suppose that the center and width for an anchor segment are μ^c and μ^w, respectively. Then the center ϕ^c and width ϕ^w of the corresponding predicted segment are therefore determined by:

$$\begin{aligned} \phi^c &= \mu^c + \alpha^c \cdot \mu^w \cdot \triangle c, \\ \phi^w &= \mu^w \cdot \exp(\alpha^w \cdot \triangle w), \end{aligned} \tag{5}$$

where α^c and α^w are both used for controlling the effect of location offsets to make location prediction stable, which are set as 0.1 empirically. As such, for a feature map with temporal dimension T_k, we can obtain a predicted segment set $\Phi_k = \{(p_j^{over}, \phi_j^c, \phi_j^w)\}_{j=1}^{T_k \cdot |R|}$. The total predicted segment set is therefore denoted as $\mathbf{\Phi} = \{\Phi_k\}_{k=1}^K$, where K is the number of temporal feature maps. We name each predicted segment as a segment-level proposal for the TSG task. As such, $\mathbf{\Phi}$ can also be regarded as the candidate proposal set.

3.4 Clip-level Actionness Prediction

Since the predicted segment-level proposals are generated from the hierarchical temporal convolutional architecture whose temporal unit spans are regularly distributed or manually defined over the video sequence, the temporal boundaries of these proposals are inevitably limited and imprecise with respect to the unconstrained sentence specified target segments. To further improve the accuracy of the temporal boundaries of the predicted segment-level proposals, we introduce clip-level actionness prediction to adjust the proposal boundaries in a fine-grained manner [7], [19].

Specifically, the clip-level actionness prediction consists of starting point, ending point and middle point probability predictions, which are realized by three sub-networks, respectively. These three sub-networks work on the first temporal feature map $\mathbf{A}_1$ and are of the same configuration without sharing weights. Each sub-network is composed of two layers of semantic modulated temporal convolution. Taking the middle point probability prediction as an example, the first layer of semantic modulated temporal convolution is set as kernel size 3, stride size 1, and number of filters d_h, yielding $\text{Conv}(3, 1, d_h)$, and the second layer is configured as $\text{Conv}(3, 1, 1)$. As such, we obtain three probability sequences $P^s = \{p_k^s\}_{k=1}^T$, $P^e = \{p_k^e\}_{k=1}^T$, $P^m = \{p_k^m\}_{k=1}^T$, with each p_k^s, p_k^e, and p_k^m denoting the probability that the k-th clip is at the starting point, ending point and middle point of the target segment, respectively.

The three probability sequences will further be used to adjust and refine the temporal boundaries of the predicted segment-level proposals in the inference stage. In addition, the clip-level actionness prediction shares the bottom temporal convolutional layer with the segment-level proposal generation, which is also expected to help the segment-level prediction. More detailed analyses will be provided in Sec. 4.5.

3.5 Training and Inference

3.5.1 Training

Since our proposed model contains both segment-level proposal prediction and clip-level actionness prediction, during the training process, both components cooperate with each other and are jointly trained in an end-to-end fashion. Specifically, the training objectives of our proposed model are composed of both segment-level objective function L_{seg} and clip-level objective function L_{clip}:

$$L_{all} = L_{seg} + L_{clip}. \tag{6}$$

Segment-level Objective. Our training sample consists of three elements: an input video, a sentence query, and the ground-truth segment. We treat anchor segments within different temporal feature maps as positive if their tIoUs (temporal Intersection-over-Union) with ground-truth segments are larger than 0.5. Our segment-level objective includes an overlap prediction loss L_{over} and a location prediction loss L_{loc}. The L_{over} term is realized as a cross-entropy loss, which is defined as:

$$\begin{aligned} L_{over} = \sum_{z \in \{pos, neg\}} -\frac{1}{N_z} \sum_{i=1}^{N_z} & g_i^{over} \log(p_i^{over}) \\ & + (1 - g_i^{over}) \log(1 - p_i^{over}), \end{aligned} \tag{7}$$

where g^{over} is the ground-truth tIoU between the anchor and target segments, and p^{over} is the predicted overlap score. The L_{loc} term measures the Smooth L_1 loss [37] for positive samples:

$$L_{loc} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \text{smooth}_{L1}(g_i^c - \phi_i^c) + \text{smooth}_{L1}(g_i^w - \phi_i^w), \tag{8}$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x|<1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{9}$$

Here g^c and g^w are the center and width of the ground-truth segment, respectively.

The two losses are jointly considered for the segment-level prediction, with λ_{over} and λ_{loc} balancing their contributions:

$$L_{seg} = \lambda_{over} L_{over} + \lambda_{loc} L_{loc}. \tag{10}$$

Clip-level Objective. Since the clip-level actionness prediction produces three probability sequences P^s, P^e, and P^m for each clip, to achieve the clip-level training, we first generate the corresponding starting/ending/middle label sequences accordingly. Given one groundtruth segment (t^s, t^e) for one sentence query in the training set, if the timestamp t^s/t^e lies in one clip, the starting/ending label of that clip will be set as 1, otherwise 0. Meanwhile, the middle labels of the clips which lie in the groundtruth segment (t^s, t^e) are set as 1, otherwise 0. In this way, we obtain the groundtruth labels for the starting/ending/middle probability sequences, which are denoted as $G^s = \{g_k^s\}_{k=1}^T$, $G^e = \{g_k^e\}_{k=1}^T$, and $G^m = \{g_k^m\}_{k=1}^T$, respectively.

Given the predicted probability sequences and groundtruth labels, we compute the cross-entropy loss between each G^x and P^x pair ($x \in \{s, e, m\}$) as defined in Eq. (7), and then get the corresponding starting loss L_{clip}^s, ending loss L_{clip}^e, and middle loss L_{clip}^m. The clip-level objective function is defined as the weighted sum of these three loss terms as follows:

$$L_{clip} = \lambda_s L_{clip}^s + \lambda_e L_{clip}^e + \lambda_m L_{clip}^m, \tag{11}$$

TABLE 1
Notations used in the inference section. The first (top) part of this table indicates the outputs of the segment-level proposal prediction, and the second part denotes the outputs of the clip-level actionness prediction, and the third part presents some notations used in temporal boundary adjustment procedure.

Symbol	Definition
$\mathbf{\Phi}$	segment-level proposal set
Φ_i	one segment-level proposal from $\mathbf{\Phi}$
ϕ_i^s/ϕ_i^e	starting/ending point of Φ_i
p_i^{over}	overlap score between Φ_i and groundtruth
P^s	starting probability sequence
P^e	ending probability sequence
P^m	middle probability sequence
ξ_i^s	relaxation interval around the starting point of Φ_i
ξ_i^e	relaxation interval around the ending point of Φ_i
$\phi_i^{s,max}$	point of maximal starting probability in ξ_i^s
$\phi_i^{e,max}$	point of maximal ending probability in ξ_i^e
Φ_i^*	proposal after adjusting the boundary of Φ_i
$\phi_i^{s,*}/\phi_i^{e,*}$	starting/ending point of Φ_i^*
$p_i^{over,*}$	overlap score between Φ_i^* and groundtruth
$\mathbf{\Phi}^b$	proposal set after boundary adjustment and NMS
$\mathbf{\Phi}^{tag}$	TAG-based proposal set with respect to P^m
$\mathbf{\Phi}^{b,tag}$	proposal set after intergrating $\mathbf{\Phi}^b$ and $\mathbf{\Phi}^{tag}$

where λ_s,λ_e, and λ_m are the weights to balance the different loss terms.

3.5.2 Inference

As we mentioned before, the predicted proposal set $\mathbf{\Phi}$ localizes different temporal segments with various scales and locations, while the temporal boundaries of these proposals are limited, yielding imprecise TSG results. The actionness prediction provides an evaluation of each video clip about whether it locates in the starting/ending/middle parts of the target segment [7], [19], which is more sensitive to the boundaries of the sentence specified locations. Hence, we propose to leverage the actionness prediction results to adjust and refine the temporal boundaries of the predicted segment-level proposals in the inference phase. To help the understanding of our inference procedure, some notations in this section are summarized in Table 1.

Temporal boundary adjustment based on the predicted starting and ending probability sequences. Specifically, suppose $\Phi_i = (\phi_i^s, \phi_i^e, p_i^{over})$ is one proposal from the predicted segment-level proposal set $\mathbf{\Phi} = \{\Phi_i\}_{i=1}^M$. Here ϕ_i^s and ϕ_i^e denote the starting and ending points of the proposal Φ_i, and p_i^{over} is the corresponding overlap score. M is the total number of the proposals in $\mathbf{\Phi}$. We will adjust ϕ_i^s and ϕ_i^e according to the predicted starting and ending probability sequences P^s and P^e, respectively. Firstly, we define two relaxation intervals for the proposal:

$$\begin{aligned} \xi_i^s &= [\phi_i^s - d_i/\varepsilon, \phi_i^s + d_i/\varepsilon], \\ \xi_i^e &= [\phi_i^e - d_i/\varepsilon, \phi_i^e + d_i/\varepsilon], \end{aligned} \tag{12}$$

where $d_i = \phi_i^e - \phi_i^s$ is the temporal length of the proposal Φ_i, and ε is the scale factor which controls the size of the relaxation intervals. We further define the temporal point which has the maximal starting probability in ξ_i^s as $\phi_i^{s,max}$, and the corresponding maximal starting probability score in P^s is defined as $p_i^{s,max}$. Accordingly, the max ending probability score and the corresponding temporal point in the interval ξ_i^e are defined as $p_i^{e,max}$ and $\phi_i^{e,max}$, respectively. If $p_i^{s,max}$ or $p_i^{e,max}$ is larger than a predefined threshold ν, the starting or the ending point of the proposal Φ_i will be adjusted as follows:

$$\begin{aligned} \phi_i^{s,*} &= \delta\phi_i^s + (1-\delta)\phi_i^{s,max}, \\ \phi_i^{e,*} &= \delta\phi_i^e + (1-\delta)\phi_i^{e,max}. \end{aligned} \tag{13}$$

Here ν and δ are both in $[0, 1]$, and will be set with respect to different datasets. In addition, if ϕ_i^s or ϕ_i^e is updated, the overlap score will be enlarged as $p_i^{over,*} = \tau p_i^{over}$ ($\tau > 1$). As such, a new boundary adjusted proposal $\Phi_i^* = (\phi_i^{s,*}, \phi_i^{e,*}, p_i^{over,*})$ is obtained.

NMS processing. After performing the above temporal boundary adjustment, non maximum suppression (NMS) is performed to rank and process the adjusted proposals according to the updated overlap scores $p^{over,*}$, and yield the boundary adjusted and suppressed proposal set $\mathbf{\Phi}^b$.

TAG-based proposal refinement guided by the predicted middle probability sequence. Since the middle probability sequence indicates the probability of each clip lying in the target segment, we follow TAG [38], and merge the consecutive clips with high middle probability into some temporal proposals as the candidate segments. After the TAG merging/grouping with respect to P^m, we get the TAG-based proposal set $\mathbf{\Phi}^{tag} = \{\Phi_i^{tag}\}_{i=1}^{M^{tag}}$, where M^{tag} is the total number of proposals in $\mathbf{\Phi}^{tag}$. Owing to the boundary sensitivity of the middle probability sequence, we consider both $\mathbf{\Phi}^b$ and $\mathbf{\Phi}^{tag}$ to produce a more comprehensive temporal grounding result. Specifically, for each proposal Φ_i^b in $\mathbf{\Phi}^b$, we compute its tIoU with all the proposals in $\mathbf{\Phi}^{tag}$. If the maximal tIoU is larger than 0.95, we will replace Φ_i^b with the corresponding proposal in $\mathbf{\Phi}^{tag}$. Finally, we obtain the updated proposal set $\mathbf{\Phi}^{b,tag}$, which considers both the segment-level proposals and the clip-level actionness prediction, and will be taken as the final temporal sentence grounding result.

4 Experiments

4.1 Datasets and Evaluation Metrics

We validate the performance of our proposed model on three public datasets for the TSG task, namely TACoS [20], Charades-STA [11], and AcitivtyNet Captions [39].

TACoS [20] consists of the videos collected from cooking scenes. The average length of videos in this dataset is 7 minutes. Following the previous methods, the same dataset split is used in this work, which includes 10146, 4589, 4083 query-segment pairs for training, validation, and testing, respectively.

Charades-STA [11] is built on the original Charades dataset [40], which focuses on videos containing indoor activities. The videos in Charades-STA are 30-second long on average. In total, there are 12408 and 3720 query-segment pairs for training and testing, respectively.

AcitivtyNet Captions [39] is built on ActivityNet v1.3 dataset [41]. The videos in this dataset cover a wide range of complex human activities. For each video, the temporal segment and caption sentence of each human event is annotated. Since the testing split is withheld for competition, we

follow the previous methods and merge the two validation subsets "val 1", "val 2" as the testing set. The numbers of query-segment pairs for training and testing split are 37421 and 34536, respectively.

For fair comparisons, we adopt **"R@n, IoU@m"** as our evaluation metrics as in previous works [11], [12], [13], [15], [42]. Specifically, "R@n, IoU@m" is defined as the percentage of the testing queries having at least one hitting retrieval (with IoU larger than m) in the top-n retrieved segments.

4.2 Implementation Details

Following the previous methods, 3D convolutional features (C3D [44] for TACoS and ActivityNet, and I3D [45] for Charades-STA) are extracted to encode videos, with each feature representing a 1-second video clip. According to the video duration statistics, the length of input video clips is set as 1024 for both ActivityNet Captions and TACoS, and 64 for Charades-STA to accommodate the temporal convolution. Longer videos are truncated, and shorter ones are padded with zero vectors. For the design of temporal feature maps, 7 layers with {64, 32, 16, 8, 4, 2, 1} temporal dimensions, 7 layers with {1024, 512, 256, 128, 64, 32, 16} temporal dimensions, and 9 layers with {1024, 512, 256, 128, 64, 32, 16, 8, 4} temporal dimensions are set for Charades-STA, TACoS, and ActivityNet Captions, respectively. All the first and second temporal feature maps will not be used for the segment-level proposal prediction, because the receptive fields of the corresponding feature units are too small to contain target activities. To save model memory footprint, the SCDM mechanism is only performed on the temporal feature maps which directly serve for segment-level proposal prediction.

For sentence encoding, we first embed each word in sentences with the Glove [47], and then employ a Bi-directional GRU to encode the word embedding sequence. As such, words in sentences are finally represented with their corresponding GRU hidden states. Hidden dimension of the sentence Bi-directional GRU, dimension of the multimodal fused features d_f, and the filter number d_h for temporal convolution operations are all set as 512 in this paper. The trade-off parameters of the segment-level objectives λ_{over} and λ_{loc} are set as 100 and 10, respectively. While the corresponding weights $\{\lambda_s, \lambda_e, \lambda_m\}$ for the clip-level objectives are all set as 10, 100 and 5 for the Charades-STA, TACoS, and ActivityNet Captions dataset, respectively. Adam optimizer is leveraged for model training, and the batch size is set as 16. The initial learning rate is set to 0.0001, and is gradually decayed over time. During the inference phase, the search spaces for hyper-parameters $\{\varepsilon, \nu, \delta\}$ are empirically determined according to their roles in boundary adjustment and the hyper-parameters are finally set as {6.0, 0.9, 0.7},{5.0, 0.7, 0.8},{5.0, 0.9, 0.1} on the Charade-STA, TACoS and ActivityNet Captions datasets, respectively. τ is set as 1.15 on all the three datasets. More detailed analyses for these hyper-parameters are provided in Sec. 4.7.

4.3 Compared Methods

We compare our proposed model with the following state-of-the-art baseline methods on the TSG task, namely **CTRL** [11]: Cross-model Temporal Regression Localizer, **ACRN** [15]: Attentive Cross-Model Retrieval Network, **TGN** [12]: Temporal Ground-Net, **MCF** [42]: Multimodal Circulant Fusion, **ACL** [24]: Activity Concepts based Localizer, **SAP** [16]: A two-stage approach based on visual concept mining, **Xu *et al.*** [25]: A two-stage method (proposal generation + proposal rerank) exploiting sentence re-construction, **MAN** [13]: Moment Alignment Network, **TripNet** [43]: An end-to-end reinforcement learning framework for grounding using gated-attention, **ProFree** [27]: Proposal-free Temporal Moment Localization using Guided Attention.

We use **SCDM$_{\mathbf{CAP}}$** to refer our proposed SCDM-coupled temporal convolutional network with both the segment-level proposal prediction and the clip-level actionness prediction, and **SCDM** indicates the model which only leverages the segment-level proposal prediction for temporal sentence grounding.

4.4 Performance Comparison and Analysis

Table 2 and Table 3 report the performance comparisons between our model and the existing methods on the aforementioned three public datasets. Overall, SCDM$_{\text{CAP}}$ achieves the highest temporal sentence grounding accuracy, demonstrating the superiority of our proposed model. Meanwhile, without clip-level actionness adjusting the temporal boundaries, the SCDM model has already outperformed other baseline methods, indicating the effectiveness of our proposed SCDM-modulated temporal convolutional network. Notably, SCDM$_{\text{CAP}}$ significantly outperforms the state-of-the-art methods with 6.3% improvements in the R@5,IoU@0.7 metric on the Charades-STA dataset. The performance improvements under the high IoU threshold demonstrate that SCDM$_{\text{CAP}}$ can generate grounded video segments of more precise boundaries. Although SCDM$_{\text{CAP}}$ achieves lower results of R@5,IoU@0.5 on the Charades-STA dataset, the reason is mainly due to the biased annotations in this dataset. For example, in Charades-STA, the annotated ground-truth segments are 10s on average while the video duration is only 30s on average. Randomly selecting one candidate segment can also achieve competing temporal grounding results. Therefore, the Recall values under higher IoUs are more stable and convincing even considering the dataset biases. For TACoS, the cooking activities take place in the same kitchen scene with some slightly varied cooking objects (*e.g.*, chopping board, knife, and bread, as shown in the second example of Fig 9). Thus, it is hard to localize such fine-grained activities. However, our proposed model still achieves the best results, except slight worse performances in R@5,IoU@0.3. As for the ActivityNet Captions dataset, the proposed SCDM and SCDM$_{\text{CAP}}$ also outperform other baseline methods, demonstrating that our proposed model can localize various complex human activities well.

The main reasons for our proposed model outperforming the competing models is three folds. First, the sentence information is fully leveraged to modulate the temporal convolution processes, so as to help correlate and compose relevant video contents over time to support the prediction for segment-level proposals. Second, the modulation procedure dynamically evolves with different video contents in the hierarchical temporal convolution architecture, and therefore characterizes the diverse sentence-video semantic interactions of different granularities. Third, besides the proposed

TABLE 2
Performance comparisons on the TACoS and Charades-STA datasets (%).

Method	TACoS				Charades-STA			
	R@1, IoU@0.3	R@1, IoU@0.5	R@5, IoU@0.3	R@5, IoU@0.5	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
CTRL (C3D) [11]	18.32	13.30	36.69	25.42	23.63	8.89	58.92	29.52
MCF (C3D) [42]	18.64	12.53	37.13	24.73	-	-	-	-
ACRN (C3D) [15]	19.52	14.62	34.97	24.88	-	-	-	-
SAP (VGG) [16]	-	18.24	-	28.11	27.42	13.36	66.37	38.15
ACL (C3D) [24]	24.17	20.01	**42.15**	30.66	30.48	12.20	64.84	35.13
TGN (C3D) [12]	21.77	18.90	39.06	31.02	-	-	-	-
Xu et al. (C3D) [25]	-	-	-	-	35.60	15.80	79.40	45.40
MAN (I3D) [13]	-	-	-	-	46.53	22.72	**86.23**	53.72
TripNet (C3D) [43]	23.95	19.17	-	-	38.29	16.07	-	-
ProFree (I3D) [27]	-	-	-	-	52.02	33.74	-	-
SCDM (*)	26.11	21.17	40.16	32.18	54.44	33.43	74.43	58.08
SCDM$_{\text{CAP}}$ (*)	**27.64**	**23.27**	40.06	**33.49**	**54.92**	**34.26**	76.50	**60.02**

Here we adopt C3D [44] features to encode videos on the TACoS dataset, and I3D [45] features on the Charades-STA dataset for fair comparisons. Video features adopted by other compared methods are indicated in brackets. VGG denotes VGG16 [46] features. **SCDM$_{\text{CAP}}$** refers our proposed SCDM-coupled temporal convolutional network with both the segment-level proposal prediction and the clip-level actionness prediction, and **SCDM** indicates the model which only leverages the segment-level proposal prediction for temporal sentence grounding.

TABLE 3
Performance comparisons on the ActivityNet Captions dataset (%).

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 (INP) [12]	45.51	28.47	-	57.32	43.33	-
Xu et al. (C3D) [25]	45.30	27.70	13.60	75.70	59.20	38.30
TripNet (C3D) [43]	48.42	32.19	13.93	-	-	-
ProFree (I3D) [27]	51.28	33.04	19.26	-	-	-
SCDM (C3D)	54.80	36.75	19.86	77.29	64.99	41.53
SCDM$_{\text{CAP}}$ (C3D)	**55.25**	**36.90**	**20.28**	**78.79**	**66.84**	**42.92**

Here INP denotes Inception-V4 [48] features. **SCDM$_{\text{CAP}}$** refers our proposed SCDM-coupled temporal convolutional network with both the segment-level proposal prediction and the clip-level actionness prediction, and **SCDM** indicates the model which only leverages the segment-level proposal prediction for temporal sentence grounding.

SCDM mechanism helps composing sentence-relevant video proposals for temporal grounding, the actionness prediction further adjusts and refines the boundaries of these predicted proposals in a more fine-grained clip-level, leading into more precise grounding results.

4.5 Ablation Studies

In this section, we perform two groups of ablation studies to examine the contributions of our proposed full model SCDM$_{\text{CAP}}$. Specifically, as shown in Table 4, we design seven ablation models in the first part to demonstrate the effectiveness of the proposed SCDM mechanism, and the other five models are used to verify the effectiveness of the clip-level actionness prediction. The ablation models to examine SCDM are introduced and compared at first.

- **PlainBN**: We only leverage the temporal convolution architecture to predict the segment-level proposals and process them with the NMS mechanism. The proposed SCDM is replaced with the plain batch normalization [34], and the clip-level actionness prediction is not considered. As such, PlainBN is only trained with the segment-level objective L_{seg}.
- **FC**: Based on PlainBN, we use one FC layer to introduce sentence information into the temporal convolutional architecture, which fuses each temporal feature unit with the global sentence representation $\bar{\mathbf{s}}$ after each temporal convolutional layer.
- **MUL**: Based on PlainBN, we use element-wise multiplication operation to introduce sentence information to the temporal convolutional architecture. Specifically, element-wise multiplication between each temporal feature unit and the global sentence representation $\bar{\mathbf{s}}$ is performed after each temporal convolutional layer.
- **SCM**: In this setting, the proposed SCDM first degrades to a Semantic Condition Modulation (SCM) mechanism, in which we use the global sentence representation $\bar{\mathbf{s}}$ to produce γ^c and β^c without dynamically changing these two modulation vectors with respect to different feature units. By replacing the plain batch normalization with SCM in the aforementioned PlainBN model, the new ablation model is obtained.
- **SCDM**: In this setting, we use SCDM to replace plain batch normalization in the PlainBN model.
- **DF**: In this setting, we apply the Dynamic Filter networks [28] on the aforementioned PlainBN model.
- **CondConv**: In this setting, we apply the Conditionally parameterized Convolutions [32] on the PlainBN model.

As shown in Table 4, by replacing SCDM with batch normalization, the performance of PlainBN degrades dramatically compared to the model SCDM. It indicates that only relying on multimodal fusion to exploit the relationship

TABLE 4
Ablation studies on the TACoS and Charades-STA datasets (%).

Method	TACoS				Charades-STA			
	R@1, IoU@0.3	R@1, IoU@0.5	R@5, IoU@0.3	R@5, IoU@0.5	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
PlainBN	16.82	14.97	32.53	26.67	47.52	26.91	69.85	49.35
FC	18.42	14.94	35.72	28.54	46.33	25.94	68.96	49.81
MUL	19.26	15.81	36.01	29.15	49.08	28.77	72.68	51.02
DF [28]	18.84	15.73	35.47	29.10	48.51	28.07	72.30	53.52
CondConv [32]	18.20	16.13	34.73	29.02	49.48	30.03	72.41	53.77
SCM	24.16	20.72	38.61	31.17	53.07	31.41	71.71	54.57
SCDM	26.11	21.17	**40.16**	32.18	54.44	33.43	74.43	58.08
CAP	26.28	22.72	37.89	31.79	42.72	23.49	69.36	43.31
SCDM_{CAP}(trainOnly)	27.49	22.33	39.92	33.30	54.36	33.81	76.21	59.42
SCDM_{CAP}(TBA)	27.61	23.17	40.01	33.42	54.82	34.15	76.53	59.99
SCDM_{CAP}(TAG)	27.53	22.94	39.99	33.38	54.52	33.80	76.23	59.42
SCDM_{CAP}	**27.64**	**23.27**	40.06	**33.49**	**54.92**	**34.26**	**76.50**	**60.02**

between video and sentence is not enough for the TSG task. The critical sentence semantics should be intensified to guide the temporal convolution procedure so as to better link the sentence-relevant video contents over time. However, roughly introducing sentence information in the temporal convolution architecture, like models MUL and FC, does not achieve satisfying results. Recall that temporal feature maps in the proposed model are already multimodal representations since the sentence information has been integrated during the multimodal fusion process. Directly coupling the global sentence representation $\bar{\mathbf{s}}$ with temporal feature units could possibly disrupt the visual correlations and temporal dependencies of the videos, which poses a negative effect on the temporal sentence grounding performance. Also, directly modifying the convolution kernels based on the sentence information in the DF and CondConv models makes the overall network vary greatly with the inputs and thus requires careful optimization tuning. Doing so will influence the model convergence and further lead to lower grounding accuracies. In contrast, the proposed SCDM mechanism modulates the temporal feature maps by manipulating their scaling and shifting parameters under the sentence guidance, which is lightweight while meticulous, and thereby achieves better results.

In addition, comparing SCM with SCDM, we can find that dynamically changing the modulation vectors γ^c and β^c with respect to different temporal feature units is beneficial, with R@5,IoU@0.7 increasing from 54.57% of the model SCM to 58.08% of the model SCDM on the Charades-STA dataset. The SCDM intensifies meaningful words and cues in sentences catering for different temporal feature units, with the motivation that different video segments may contain diverse visual contents and express different semantic meanings. Establishing the semantic interactions between these two modalities in a dynamic way can better align the semantics between sentence and diverse video contents, yielding more accurate grounding results.

Then, we introduce and compare the ablation models to verify the effectiveness of clip-level actionness prediction as follows:

- **CAP**: In this setting, we only use clip-level actionness prediction module to tackle the TSG task. Specifically, we first generate TAG-based proposals $\mathbf{\Phi}^{tag}$ based on the middle probability sequence. Then, we adjust the boundaries of the proposals in $\mathbf{\Phi}^{tag}$ with the starting and ending probability sequences. Each adjusted proposal will get a confidence score of $p_i^{CAP} = p_i^{s,max} \times p_i^{e,max} \times p_i^{tag}$, where $p_i^{s,max}$ and $p_i^{e,max}$ are defined in Sec. 3.5.2, and p_i^{tag} is the confidence score of the proposal referring to the TAG mechanism. The boundary adjusted proposals will be processed and ranked with the NMS mechanism based on the p^{CAP} scores. Since the segment-level proposal prediction is not adopted in this ablation model, the model is only trained with the clip-level objective L_{clip}.
- **SCDM_{CAP}(trainOnly)**: The overall model is trained with both L_{seg} and L_{clip} loss terms. However, the temporal boundary adjustment and TAG-based proposal refinement during the inference stage, as introduced in Sec. 3.5.2, are not performed. The NMS processed segment-level proposals are directly regarded as the final temporal grounding results.
- **SCDM_{CAP}(TBA)**: Based on **SCDM_{CAP}(trainOnly)**, the **SCDM_{CAP}(TBA)** model further performs the temporal boundary adjustment as stated in Sec. 3.5.2.
- **SCDM_{CAP}(TAG)**: Based on **SCDM_{CAP}(trainOnly)**, the **SCDM_{CAP}(TAG)** model further performs the TAG-based proposal refinement as stated in Sec. 3.5.2.
- **SCDM_{CAP}**: Our proposed full model, which trained with both L_{seg} and L_{clip} loss term. During the inference stage, both TBA and TAG are introduced as stated in Sec. 3.5.2. Please note that the commonly-used NMS processing is always performed in all the ablation models.

For the performance of CAP, it achieves only slightly lower performances than the SCDM-coupled temporal convolutional network on the TACoS dataset, which shows that the clip-level actionness prediction well captures the sentence-video correlation over time, and can provide good supports for the temporal boundary adjustment in a fine-grained manner. By incorporating clip-level objective in the training procedure of SCDM_{CAP}(trainOnly), the performance outperforms the SCDM model on both the Charades-STA and TACoS dataset. It demonstrates that the clip-level objective is also beneficial to the segment-level proposal prediction. Since the clip-level actionness prediction is performed on

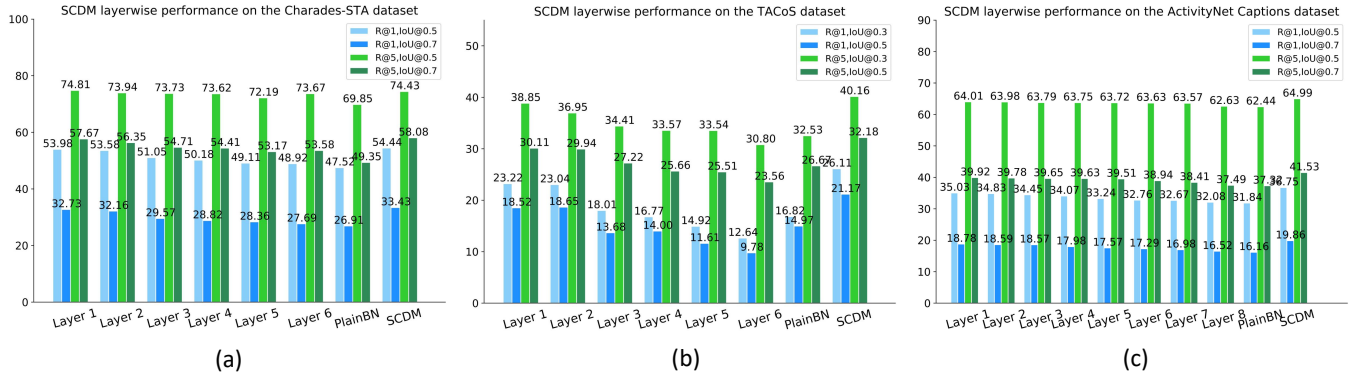


Fig. 5. Temporal sentence grounding results when SCDM is performed on different layers of the temporal convolutional architecture. Layer-x means that we only add the SCDM mechanism on the x-th layer of the temporal feature map, and Layer-1 means adding SCDM on the lowest/first layer.

the bottom layer of the temporal convolutional architecture, the actionness supervision applied on the bottom layer will guide the semantic correlation establishment on the finer clip level, and gradually help the higher layer segment proposal compositions. Based on the SCDM_{CAP}(trainOnly) model, we observe that considering either a temporal boundary adjustment based on the starting and ending probabilities in SCDM_{CAP}(TBA) or considering the TAG-based proposal refinement by merging the middle probability in SCDM_{CAP}(TAG) can further improve the temporal grounding accuracy. Combining both of the two terms, our proposed full model SCDM_{CAP} achieves the best results, which demonstrates the effectiveness of refining temporal boundaries according to the clip-level actionness scores in the inference procedure. Overall, the above ablation studies verify our design of integrating both the segment-level proposal prediction and the clip-level actionness prediction to improve the TSG performance.

4.6 SCDM Investigation for Temporal Sentence Grounding in Videos

To further investigate the effect of the proposed SCDM mechanism for the TSG task, we conduct several groups of experiments as follows.

4.6.1 Performance of SCDM on Different Layers

First, we conduct experiments by coupling SCDM on different temporal convolutional layers , and then test the performances of the yielding models on the Charades-STA, TACoS, and ActivityNet Captions datasets. The corresponding results are illustrated in Fig. 5. Specifically, since there are 7, 7, and 9 temporal feature maps in the convolution architecture on the Charades-STA, TACoS, and ActivityNet Captions datasets, respectively, there are accordingly 6, 6, and 8 layers of temporal convolutions because the first feature map are the initial multimodal fusion feature.

For all the three datasets, we can observe that when coupling SCDM on higher temporal convolution layers, the models perform worse than those coupling SCDM on lower layers. And these single-layer SCDM-coupled models achieve lower results than the model which has SCDM processing all the temporal convolution layers (please refer the last group of bars in each subfigure). Such results are evident because when SCDM guides the sentence related video content composition on the lower layer, not only the corresponding segment-level proposal prediction on this layer will be enhanced, the proposal prediction on higher layers will also be benefited through the gradually temporal convolution process. However, when coupling SCDM on higher layers, the lower layers will not be explicitly influenced except the backward gradients. With SCDM coupling in all the temporal convolutional layers, the predictions of segment proposals with various scales will be improved, and therefore produces the best results.

Moreover, the PlainBN model, which has no SCDM coupled in the convolutional architecture, achieves the worst performances in Fig. 5(a) and Fig. 5(c), while it performs better than the "*Layer4*", "*Layer5*" and "*Layer6*" models in Fig. 5(b). Such results show that the SCDM operation can improve the TSG accuracy in each temporal scale of our proposed architecture on the Charades-STA and ActivityNet Captions datasets. While on the TAcoS dataset, coupling SCDM on the higher layers may even hurt the performance. The reason may attribute to that the annotated groundtruth segments in TACoS dataset are rather small, with most of them ranging from 1s-5s length. The lowest three layers are served for grounding shorter video segments in videos, and therefore coupling SCDM on these layers is beneficial to the model performance on the TACoS dataset. However, the highest three layers on the TAcoS datasets support the temporal grounding of segments longer than 16s, which do not match the groundtruth annotation distribution very well.

4.6.2 The Influence of SCDM on Model Learning

In this subsection, we compare the training and testing performances of the SCDM-coupled model against other ablation models, to further investigate how does SCDM help learning a more effective model for the TSG task.

Fig. 6(a) shows the loss curves of different methods during the model training on the Charades-STA dataset, while Fig. 6(b) and Fig. 6(c) show the corresponding testing accuracies R@1,IoU@0.5 and R@1, IoU@0.7, respectively. Specifically, 4 models, namely PlainBN, FC, MUL, and SCDM as stated in Sec. 4.5, are compared. From the training loss and testing accuracy curves of different models, it can be

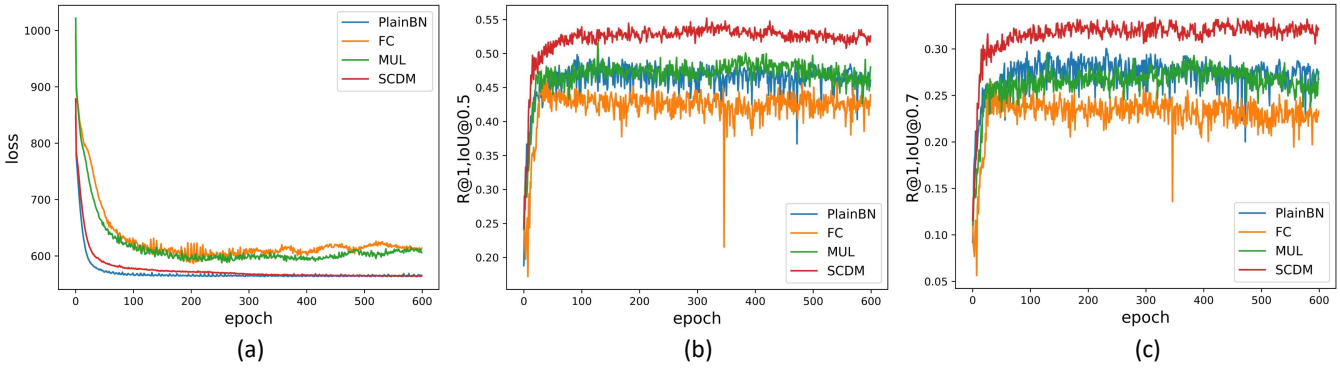


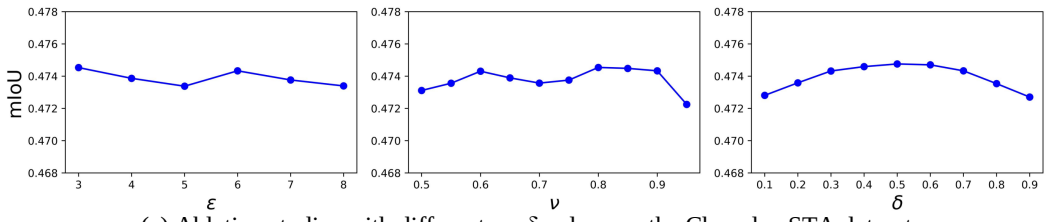
Fig. 6. Comparison of (a) training (optimization) and (b) testing (generalization) performances on the Charades-STA dataset. (a) shows the training loss curves of 4 different models PlainBN, FC, MUL and SCDM. (b) and (c) illustrate the corresponding testing accuracies R@1,IoU@0.5, R@1,IoU@0.7 of these models after each training epoch, respectively.

Fig. 7. t-SNE projections of temporal feature maps yielded by the models PlainBN and SCDM. Each temporal feature unit within these feature maps is represented by its corresponding video clip in the original video. Video clips marked with red color are within ground-truth video segments.

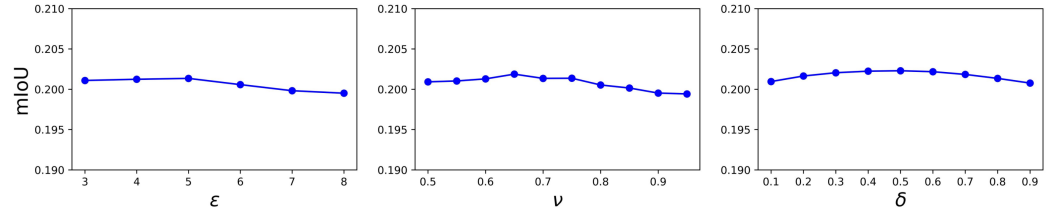
observed that compared to the model FC and MUL, the proposed SCDM-coupled model presents prominently better convergence and higher accuracy, and the loss curve of SCDM is also smoother. The loss of PlainBN decreases faster than SCDM during the initial training procedure, while these two models converge to similar loss values in the subsequent training epochs. However, the prediction accuracies of PlainBN are much lower than SCDM as shown in Fig. 6 (b) and (c). Overall, the results show that our proposed SCDM-coupled model can learn steadily and effectively from the training data, and generalize well on the testing data.

The FC, MUL, and SCDM-coupled models try to introduce sentence information to guide the temporal convolution operation for better linking sentence-relevant video contents. However, for the hierarchical temporal convolutional architecture which naturally convolves the video features through time, the sentence representation is external intervention that should be cautiously dealt with. Roughly fusing the sentence representation with the temporal feature maps, like FC and MUL, will change and disrupt the feature distribution learned from the overall hierarchical convolutional architecture heavily, and make the model more difficult to converge, as shown in Fig. 6 (a). Accordingly, such under-fitting models will inevitably yield lower temporal grounding accuracies, as illustrated in Fig. 6 (b) and (c).

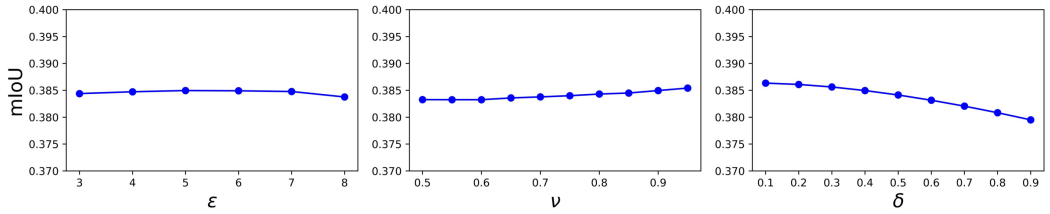
Our proposed SCDM does not directly modify the temporal feature maps, but instead controls the feature normalization parameters with the sentence semantics, and therefore meticulously shifts and scales the feature map before it passes to the next temporal convolution. Since the feature normalization mechanism like Batch Normalization [34] is a proven technique that can improve the training of neural networks by stabilizing the distributions of layer inputs [49], our proposed SCDM modulates and stabilizes the distribution of temporal feature maps with the guidance of sentence semantics, and therefore achieves good model convergence. Compared to PlainBN, our SCDM-coupled model converges slower during the initial training procedure. The reason is due to that the feature normalization parameters of PlainBN are purely learned from the temporal convolutional architecture, and they can well fit the training data with the backbone model. Our SCDM explicitly generates the normalization parameters from the external sentence representations, which brings additional knowledge or intervention for the TSG training procedure, and thereby increases the data fitting complexity and slows the model convergence. However, SCDM also brings advantages that the



(a) Ablation studies with different ε,υ,δ values on the Charades-STA dataset



(b) Ablation studies with different ε,υ,δ values on the TACoS dataset



(c) Ablation studies with different ε,υ,δ values on the ActivityNet Captions dataset

Fig. 8. Temporal grounding accuracies with different ε,ν,δ values on the Charades-STA, TACoS, and ActivityNet Caption datasets.

explicit sentence-guided feature modulation leads the model to better correlate the semantic-coherent video contents, and the increased data-fitting complexity from the other hand alleviates the over-fitting problem, and therefore makes the SCDM-coupled model achieve much higher testing accuracy.

4.6.3 The Comparison of Temporal Feature Maps w/wo SCDM

To investigate whether SCDM can link the sentence-relevant video contents in semantics, we visualize the features units in temporal feature maps produced by the model PlainBN and the model SCDM in Fig. 7. For both of the trained models, we extract their temporal feature maps, and subsequently apply t-SNE [50] to each temporal feature unit within these maps. Since each temporal feature unit corresponds to one specific location in the original video, we then assign the corresponding video clips to the positions of these feature units in the t-SNE embedded space. As illustrated in Fig 7, temporal feature maps of 2 testing videos are visualized, where the video clips marked with red color denote the ground-truth segments of the given sentence queries. Interestingly, it can be observed that through SCDM processing, video clips within ground-truth segments are more tightly grouped together. In contrast, the clips without SCDM processing are separated in the learned feature space. This demonstrates that SCDM successfully associates the sentence-relevant video contents according to the sentence semantics, which is beneficial to the later segment-level proposal prediction.

4.7 Analysis on Hyper-parameters for Temporal Boundary Adjustment

As stated in Sec. 3.5.2, the hyper-parameters ε, ν, and δ directly influence the temporal grounding results. In this subsection, we introduce the hyper-parameter choosing procedure, and test the model sensitivity with respect to their values.

We determine the ε, ν and δ values by first empirically setting some initial search spaces according to their roles in our temporal boundary adjustment procedure, and then search the specific parameters that lead to the best results on the Charades-STA, TACoS, and ActivityNet Captions datasets.

For the ε value, according to Equation (12) in Sec. 3.5.2, it is used to control the length of the relaxation intervals ξ_i^s and ξ_i^e, and the larger ε makes the intervals shorter. Meanwhile, if $\varepsilon < 2$, ξ_i^e will intersect ξ_i^s, making the adjusted ending point be possibly before the starting point, which results in invalid temporal segments. Therefore, the ε values should be larger than 2. In order to set valid and relatively broad intervals for our temporal boundary adjustment around the starting and ending points, the search space for ε is set as $\{3, 4, 5, 6, 7, 8\}$. For the ν value, if the starting/ending actionness scores are larger than it, the temporal boundary adjustment will be performed. Therefore, to ensure the confidence and accuracy of our boundary adjustment, we set $0.5 \leq \nu < 1.0$, yielding the search space $\{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$. The δ value determines how to move the starting/ending points of the proposal to obtain the final result. Smaller δ means moving the starting/ending

points of the proposal towards the points with maximum starting/ending actionness scores. We directly set the search space of δ as $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

We conduct experiments on the Charades-STA, TACoS and ActivityNet Captions datasets to search and tune the ε, ν, and δ values. Note that Charades-STA does not have a validation set. ActivityNet Captions does not publish the testing set, with the validation set used as the testing set for researchers to evaluate their results. We randomly choose 10% samples in the training sets of the Charades-STA and ActivityNet Captions datasets as their corresponding validation sets to tune the parameters. For the TACoS dataset, we directly leverage their provided validation set. By greedily searching the ε, ν, and δ values, we find that the TSG performance is insensitive to the choice of these hyper-parameters. For reference, we plot the mIoU (We compute the IoU between the rank1-predicted segment and the groundtruth for each video-sentence pair, and mIoU means the average IoU between all of the validation pairs) of our proposed model when fixing two parameters in their best values and varying the rest one. The results on the three datasets are shown in Fig. 8. We can clearly see that the performance variations with different ε, ν, and δ values are very small, which verifies that our proposed model is indeed insensitive to these hyper-parameters.

4.8 Model Efficiency Comparisons

Table 5 illustrates the run-time efficiency, model size (#param), and memory footprint during the training procedure of different methods. Specifically, "Run-Time" denotes the average time to localize one sentence in a given video during the inference stage. The methods with released codes are tested with one Nvidia TITAN XP GPU. The experiments are performed on the TACoS dataset since the videos in this dataset are relatively long (7 minutes on average), and are appropriate to evaluate the temporal grounding efficiency of different methods. It can be observed that our SCDM-coupled temporal convolution architecture without clip-level actionness prediction module (the third row) achieves the fastest run-time with the smallest model size. Both CTRL and ACRN methods need to sample candidate segments with various sliding windows in the videos first, and then match the input sentence with each segment individually. Such a two-stage architecture will inevitably influence the temporal sentence grounding efficiency, since the matching procedure through sliding window is quite time-consuming. In contrast, our proposed model adopts a hierarchical convolution architecture, and naturally covers multi-scale video segments for grounding with multi-layer temporal feature maps. Thus, we only need to process the video in one pass of temporal convolution and then yield the TSG results, and achieve higher efficiency. In addition, SCDM only needs to control the feature normalization parameters and is lightweight towards the overall convolution architecture. Therefore, our SCDM coupled model also has smaller model size.

By incorporating the clip-level actionness prediction module, the model size and the memory footprint during the training stage of the models SCDM_{CAP}(trainOnly) and SCDM_{CAP} increase, as shown in the fourth and fifth row in Table 5. Meanwhile, comparing the fourth row to the third tow, the average run-time to localize a sentence for the SCDM_{CAP}(trainOnly) model is similar to the purely segment-level prediction model SCDM. It is evident that we only use the clip-level objective for the SCDM_{CAP}(trainOnly) model training while do not perform the temporal boundary adjustment in the inference stage, and the SCDM_{CAP}(trainOnly) inference procedure is exactly the same with that of the SCDM model. However, if the temporal boundary adjustment is performed, the run-time for the full model SCDM_{CAP} is larger. Such result is due to that the boundary adjustment will search through the actionness score sequences, and the TAG grouping also needs extra NMS operation, which are both time-consuming. Since the SCDM and SCDM_{CAP}(trainOnly) models also achieve good performances as illustrated in Table 4, they can be applied for the TSG task if the model efficiency needs to be considered in some scenarios.

TABLE 5
Comparison of model running efficiency, model size, and memory footprint.

Method	Run-Time	Model Size	Memory Footprint
CTRL [11]	2.23s	22M	725MB
ACRN [15]	4.31s	128M	8537MB
SCDM	0.78s	15M	4533MB
SCDM_{CAP}(trainOnly)	0.80s	18M	7467MB
SCDM_{CAP}	1.48s	18M	7467MB

4.9 Qualitative Results

Some qualitative examples of our model are illustrated in Fig 9. Evidently, our model can produce accurate segment boundaries for the TSG task. Moreover, we also visualize the attention weights (defined in Eq. (2)) produced by SCDM when it processes different temporal units. It can be observed that different video contents attentively trigger different words in sentences. For example, the "*walking & open*" in case (a) and the "*returns*" in case (b) get higher attention weights in the target temporal regions. Those words are indeed important cues for identifying the target video segments, and intensifying their semantics in the SCDM procedure can better recognize and link the target-video contents over time and thus benefit the subsequent temporal proposal predictions. To further demonstrate the benefits of our proposed clip-level actionness prediction module, we also show some temporal grounding results before and after the boundary adjustment and refinement, as well as the predicted starting, middle and ending probability scores in Fig. 10. In these illustrated examples, the adjusted temporal boundaries of the predicted segments with pink color are more precise than the segment-level proposal predictions with brown color. Even in the case (b), the final predicted segment is exactly the groundtruth segment. Meanwhile, by examining the predicted actionness scores, we can also find that the higher scored regions indeed correspond to the starting/ending/happening of specific activities in videos. Such results show the clip-level actionness prediction can well capture the activity evolution in videos in a fine-grained level, based on which the temporal boundaries of the predicted segments can therefore be adjusted to more precise positions.

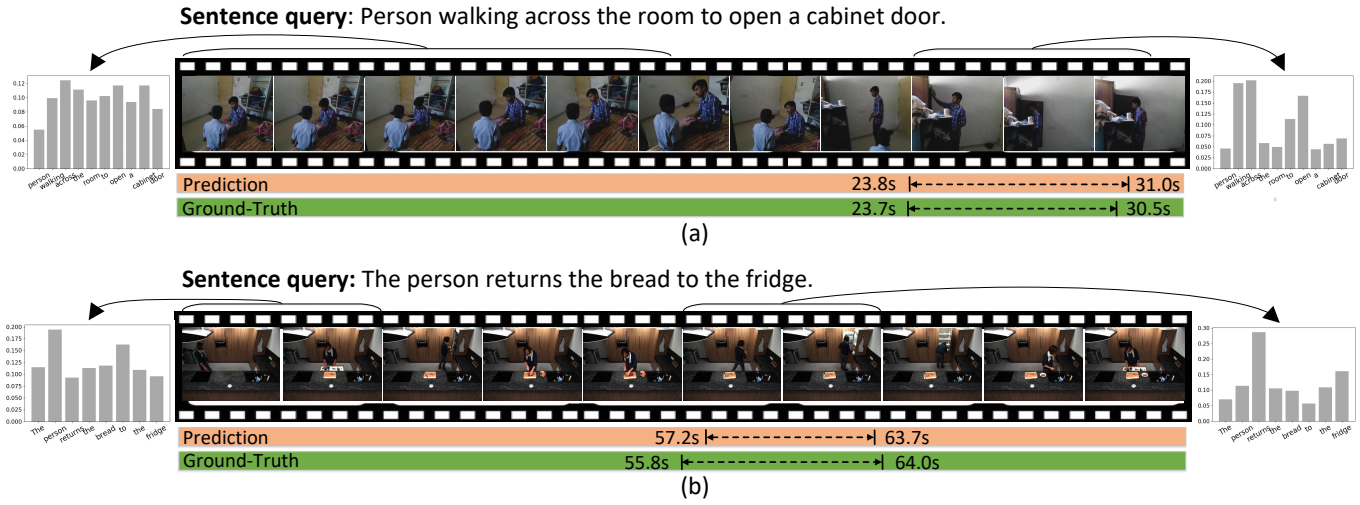


Fig. 9. Qualitative prediction examples of our proposed model for temporal sentence grounding in videos. The rows with green background show the ground-truths for the given sentence queries, and the rows with pink background show the final location prediction results. The gray histograms show the word attention weights produced by SCDM at different temporal regions.

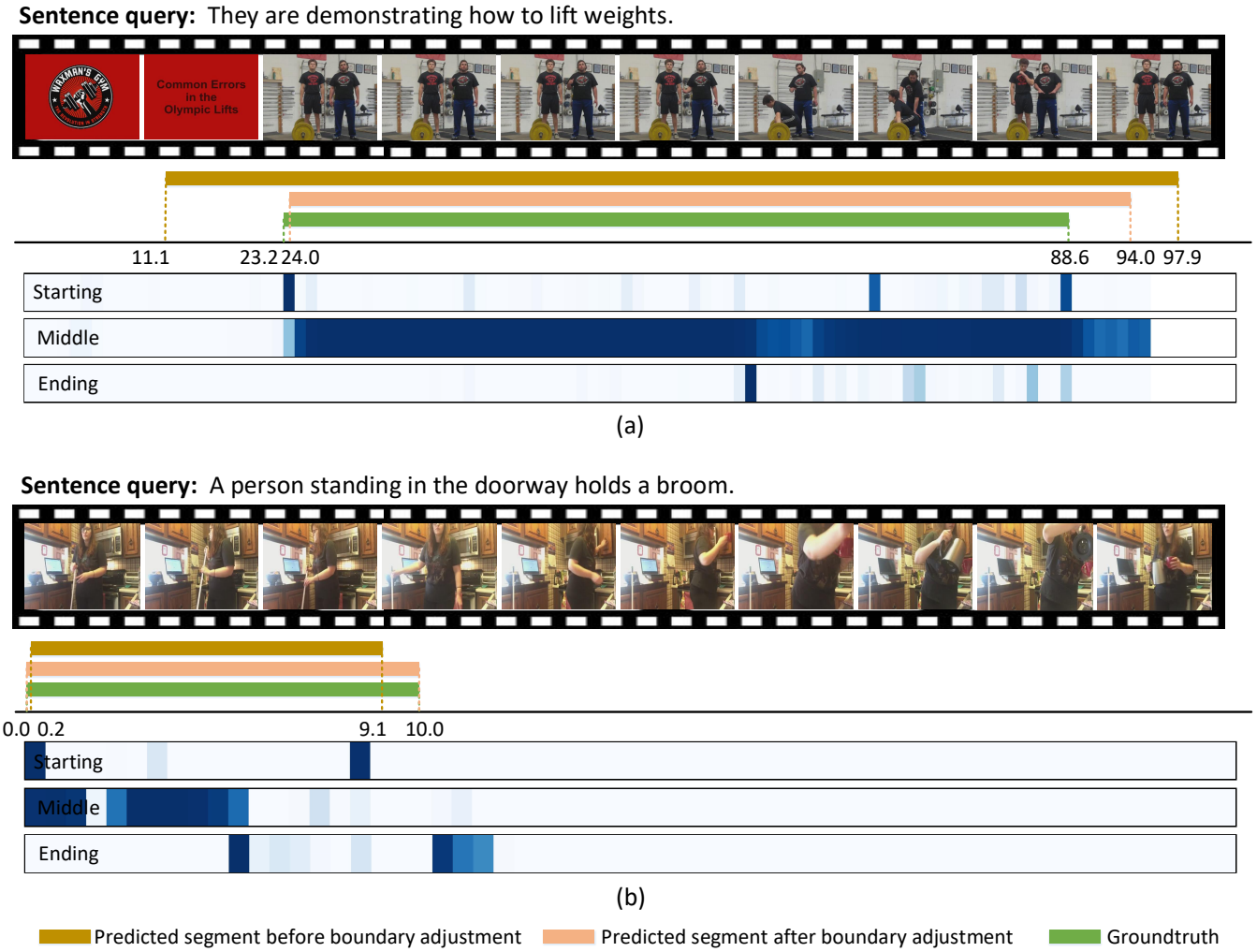


Fig. 10. Qualitative prediction examples of our proposed model with the clip-level actionness prediction. The rows with green background show the groundtruths for the given sentence queries, the rows with brown and pink background show the predicted segments before and after the temporal boundary adjustment and refinement, respectively. The bottom three heatmaps show the predicted starting, middle and ending probability sequences, respectively. The deeper the color, the higher the probability score.

5 Conclusions

In this paper, we proposed a novel semantic conditioned dynamic modulation (SCDM) mechanism for tackling the TSG task. The proposed SCDM leverages the sentence semantics to modulate the temporal convolution operations to better correlate and compose the sentence-relevant video contents over time. Coupling the SCDM mechanism with a hierarchical temporal convolutional arcthiecture, the proposed SCDM dynamically evolves with the diverse video contents of different temporal granularities, and the sentence described video segments are therefore tightly correlated and composed, leading to more accurate temporal boundary predictions. Meanwhile, the SCDM modulated temporal convolutions were further exploited to predict actionness scores for video sequences, which are used to adjust the temporal boundaries of the predicted video segments and further improve the temporal grounding accuracy. The experimental results obtained on three widely-used datasets demonstrate the superiority of the SCDM-coupled temporal convolutional architecture on the TSG task.

Acknowledgments

This work was supported by National Natural Science Foundation of China Major Project No.U1611461 and Shenzhen Nanshan District Ling-Hang Team Grant under No.LHTD20170005. Yitian Yuan is partially supported by the Tencent Elite Internship Program.

References

[1] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 988–996.

[2] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognition Challenge*, vol. 1, no. 2, p. 2, 2014.

[3] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[4] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3093–3102.

[5] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. M. Snoek, "Actor and action video segmentation from a sentence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5958–5966.

[6] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Lecun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

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

[8] Y. Feng, L. Ma, W. Liu, T. Zhang, and J. Luo, "Video re-localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 51–66.

[9] Y. Feng, L. Ma, W. Liu, and J. Luo, "Spatio-temporal video re-localization by warp lstm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1288–1297.

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

[11] 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.

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

[13] 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.

[14] J. Chen, L. Ma, X. Chen, Z. Jie, and J. Luo, "Localizing natural language in videos," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

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

[16] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *AAAI Conference on Artificial Intelligence*. IEEE, 2019.

[17] Z. Chen, L. Ma, W. Luo, and K.-Y. K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," in *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[18] Y. Yuan, L. Ma, and W. Zhu, "Sentence specified dynamic video thumbnail generation," in *27th ACM International Conference on Multimedia.*, 2019.

[19] 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. 3–19.

[20] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013.

[21] I. Naim, Y. C. Song, Q. Liu, H. A. Kautz, J. Luo, and D. Gildea, "Unsupervised alignment of natural language instructions with video segments." in *AAAI Conference on Artificial Intelligence*, 2014, pp. 1558–1564.

[22] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "Book2movie: Aligning video scenes with book chapters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1827–1835.

[23] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid, "Weakly-supervised alignment of video with text," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4462–4470.

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

[25] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *AAAI Conference on Artificial Intelligence*, vol. 2, no. 6, 2019, p. 7.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations*, 2017.

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

[28] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in neural information processing systems*, 2016, pp. 667–675.

[29] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 30–38.

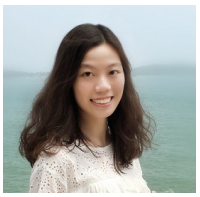
[30] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," in *International Conference on Learning Representations, ICLR*, 2017.

[31] E. A. Platanios, M. Sachan, G. Neubig, and T. Mitchell, "Contextual parameter generation for universal neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 425–435.

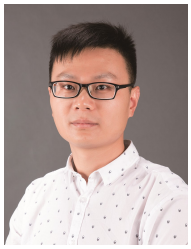
[32] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," in *Advances in Neural Information Processing Systems*, 2019, pp. 1307–1318.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *European Conference on Computer Vision*, pp. 21–37, 2016.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 448–456.
[35] H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," *Neural Information Processing Systems*, pp. 6594–6604, 2017.
[36] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *International Conference on Learning Representations*, 2017.
[37] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
[38] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection." *arXiv preprint arXiv:1703.02716*, 2017.
[39] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 706–715.
[40] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," *European Conference on Computer Vision*, pp. 510–526, 2016.
[41] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.
[42] A. Wu and Y. Han, "Multi-modal circulant fusion for video-to-language and backward," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 1029–1035.
[43] M. Hahn, A. Kadav, J. M. Rehg, and H. P. Graf, "Tripping through time: Efficient localization of activities in videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
[44] 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.
[45] 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.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.
[47] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.
[48] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[49] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization," in *NIPS'18 Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2488–2498.
[50] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.



Yitian Yuan is currently a Ph.D. student in the Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University. Yitian received her B.E. degree from the Department of Computer Science and Technology of Beijing Jiaotong University in 2016. Her main research interests include multimedia analysis, computer vision and deep learning.



Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noah'Ark Laboratory, Hong Kong, from 2013 to 2016. He is currently a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China. His current research interests lie in the areas of computer vision, multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment.

Dr. Ma received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a finalist in HKIS Young Scientist Award in engineering science in 2012.

Jingwen Wang received the Ph.D. degree in computer science and technology from South China University of Technology, Guangzhou, China, in 2018. He is currently a senior researcher with Tencent AI Lab, Shenzhen, China. His research interests include deep learning and computer vision, with respect to action classification, action detection, vision and language.

Wei Liu (M'14-SM'19) is currently a Distinguished Scientist of Tencent, China and a director of Computer Vision Center at Tencent AI Lab. Prior to that, he has been a research staff member of IBM T. J. Watson Research Center, Yorktown Heights, NY, USA from 2012 to 2015. Dr. Liu has long been devoted to research and development in the fields of machine learning, computer vision, pattern recognition, information retrieval, big data, etc. Dr. Liu currently serves on the editorial boards of IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition, etc. He is a Fellow of the International Association for Pattern Recognition (IAPR) and an Elected Member of the International Statistical Institute (ISI).

Wenwu Zhu is currently a Professor and Deputy Head of Computer Science Department of Tsinghua University and Vice Dean of National Research Center on Information Science and Technology. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs New Jersey as a Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996.

He served as the Editor-in-Chief for the IEEE Transactions on Multimedia (T-MM) from January 1, 2017 to December 31, 2019. He has been serving as Vice EiC for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) and the chair of the steering committee for IEEE T-MM since January 1, 2020. His current research interests are in the areas of multimedia computing and networking and big data. He has published over 400 papers in the referred journals and received ten Best Paper Awards including IEEE TCSVT in 2001 and 2019, and ACM Multimedia 2012. He is an IEEE Fellow, AAAS Fellow, SPIE Fellow and a member of the European Academy of Sciences (Academia European).