Can Shuffling Video Benefit Temporal Bias Problem: A Novel Training Framework for Temporal Grounding

Jiachang Hao, Haifeng Sun*, Pengfei Ren, Jingyu Wang*, Qi Qi, and Jianxin Liao

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{haojc, hfsun, rpf, wangjingyu, qiqi8266}@bupt.edu.cn jxlbupt@gmail.com

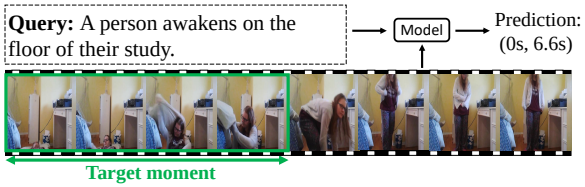
Abstract. Temporal grounding aims to locate a target video moment that semantically corresponds to the given sentence query in an untrimmed video. However, recent works find that existing methods suffer a severe temporal bias problem. These methods do not reason the target moment locations based on the visual-textual semantic alignment but over-rely on the temporal biases of queries in training sets. To this end, this paper proposes a novel training framework for grounding models to use shuffled videos to address temporal bias problem without losing grounding accuracy. Our framework introduces two auxiliary tasks, cross-modal matching and temporal order discrimination, to promote the grounding model training. The cross-modal matching task leverages the content consistency between shuffled and original videos to force the grounding model to mine visual contents to semantically match queries. The temporal order discrimination task leverages the difference in temporal order to strengthen the understanding of long-term temporal contexts. Extensive experiments on Charades-STA and ActivityNet Captions demonstrate the effectiveness of our method for mitigating the reliance on temporal biases and strengthening the model's generalization ability against the different temporal distributions. Code is available at https://github.com/haojc/ShufflingVideosForTSG.

Keywords: Temporal Grounding; Temporal Bias; Video and Language;

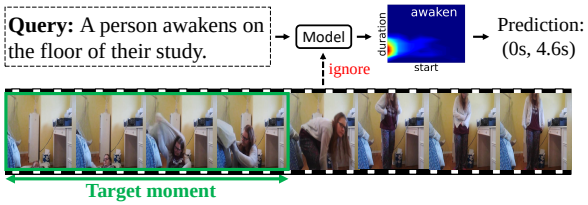
1 Introduction

Temporal grounding [11,16] aims to localize the relevant video moment of interest semantically corresponding to the given sentence query in an untrimmed video, as illustrated in Fig. 1a. Due to its vast potential applications in video captioning, video question answering, and video retrieval, this task has attracted increasing interest over the last few years. However, this task suffers a temporal bias problem [37,52,51], which severely hinders the development of the temporal grounding task.

* Corresponding author



(a) **Temporal grounding task**



(b) **Temporal bias problem**

Fig. 1: **(a)** Temporal grounding is to localize a moment with the start point (0s) and end point (6.6s) in the video for the query. **(b)** An example of temporal bias problem: a model ignores the visual input and uses the memorized temporal bias of the word 'awaken' in the training set of Charades-STA to make the prediction.

Temporal bias problem refers to that a method reasons the target moment locations not based on the visual-textual semantic matching but over-relies on the temporal biases of queries in training set [37]. As illustrated in Fig. 1b, a grounding model ignores visual inputs and takes a shortcut to directly exploit the memorized temporal biases of the given query in training set to reason the location of the target video moment. The temporal bias problem severely hinders the development of temporal grounding [37,51]. Because when we give a sentence query, we wish the target moment to be localized based on the query semantics wherever the target moment is. [37,52] found that many state-of-the-art methods [11,54,60,53,34,58,55] suffer this problem and perform poor generalization ability against the different temporal distributions. Some methods [53,60] even do not make any use of the visual input during reasoning. To this end, this paper aims to mitigate the excessive reliance on temporal biases and strengthen the model's generalization ability against the different temporal distributions.

Video-query pairs in existing datasets have high correlations between queries and ground-truth temporal positions of target moments, making it possible for a grounding model to take a shortcut [37,51]. So we ask *whether we can shuffle videos to break these correlations to address the temporal bias problem.* We can shift randomly target moments to other temporal positions in videos to dilute the temporal bias of the corresponding query in training set. Thus the shortcut of memorizing temporal biases will be ineffective, and the model has to turn the attention back to the visual contents semantically matching queries. However, directly using these shuffled videos as training samples is not appropriate. On

the one hand, these shuffled videos may not match the real situation, causing a poor generalization ability of the grounding model. On the other hand, shuffling videos disturbs the long-term temporal contexts within videos. These incorrect contexts may weaken the perception ability of grounding models for long-term temporal relations, which is important for temporal grounding [60,17].

To this end, instead of using the shuffled videos as augmented training samples, we propose to take the shuffled and original videos as paired input and design auxiliary tasks to promote the grounding model training. We design two auxiliary tasks, cross-modal matching and temporal order discrimination, to mine the cross-modal semantic relevance from the paired videos and mitigate the reliance on temporal biases. The cross-modal matching task requires that the model predicts as consistent frame-level cross-modal relevance as possible for target moments, even if their temporal positions change. This task encourages the grounding model to focus on spatial and short-term visual contents to semantically match queries. The temporal order discrimination task is to discriminate whether the video moment sequence is in correct temporal order. This task guides the grounding model to learn the correct temporal contexts and thus strengthens the perception ability on long-term temporal relations.

We propose a span-based framework to handle the temporal grounding and two auxiliary tasks. And our method can be easily transferred to other grounding models. To sum up, the main contributions of our work are as follows:

(1) We propose a novel training framework for temporal grounding models to use shuffled videos to address the temporal bias problem.

(2) Our cross-modal matching task can mitigate the model's reliance on temporal biases and turn the attention back to the visual-textual semantic matching, and the temporal order discrimination task can strengthen the understanding of long-term temporal relations.

(3) Extensive experiments on Charades-STA and ActivityNet Captions demonstrate the effectiveness of our method for mitigating the grounding model's reliance on temporal biases and strengthening the generalization ability against the different temporal distributions. And we achieve state-of-the-art on the redivided splits for temporal bias problem.

2 Related Work

Temporal Grounding Temporal grounding, also known as temporal sentence grounding in video and video moment retrieval, was first proposed by [11,16]. Proposal-based methods formulate this task as a ranking task to find the best matching video proposal for a given sentence query. These methods first generate the candidate video segments by slide windows [11,31,30] or a proposal network [50,6,48,25] or predefined anchors [3,23,56,62,22,27,26,28,1,29,24], and then semantically match each candidate with the sentence query. However, the proposal generation and semantic matching for all the proposals are resource-consuming and inefficient. To discard proposals, proposal-free methods encode the video modality only once and directly interact each video frame with the

sentence query. Specifically, regression-based methods [54,32,5,7,8] regress the temporal coordinates of the localized video moment from a compact representation. Span-based methods [13,39,4,61,18,35,57,14] predict the probabilities of each frame being the start/end of the location.

Temporal Bias Problem [37] first proposed the temporal bias problem. They found that the popular datasets for temporal grounding task include significant biases through explicit statistics about temporal locations of the top-50 verbs in datasets. Then they verified that some state-of-the-art models did not achieve cross-modal alignment but exploited dataset biases instead. To correctly evaluate grounding performance, [52] re-divided the splits of two popular datasets, Charades-STA and ActivityNet Captions, to make the training and test sets have different temporal distributions of queries. To address the temporal bias problem, DCM [51] disentangles temporal position information from each proposal feature via a constraint loss and then leverages causal intervention to fairly consider all candidate proposals. To evaluate the effectiveness, DCM [51] simulated the out-of-distribution test samples by inserting a sequence of generated video features at the beginning of the original video feature sequence. However, this simulation is not much convincing. On the one hand, the inserted features are generated from a normal distribution so that these features may not contain any meaningful content. On the other hand, the length of the inserted video sequence is too short to change the temporal biases significantly. In this paper, we use the re-divided splits [52] to evaluate performance and generalization ability.

Temporal Relation Modeling Temporal relation modeling is a fundamental problem in video understanding. Typical methods apply RNN [36], 3D-CNN [19,2,42] to capture the short-term temporal relations and non-local [47], Transformer [43], TDN [46] to capture the long-term ones. However, due to the black box of neural networks, it is uncertain whether the features learned from the aforementioned mechanism contain the temporal relations or other information like scene, appearance, and temporal bias. Therefore, some works [10,33,9,21,49] attempt to explicitly strengthen video representation learning in temporal relations by using auxiliary tasks. Temporal order verification is one of the most frequently used tasks because it does not require extra annotations. This task aims to determine whether a sequence of frames from a video is in the correct temporal order. Inspired by that, we design two auxiliary tasks for temporal grounding models to promote the visual-textual matching features mining.

3 Methods

3.1 Problem Formulation

Given an untrimmed video V and a sentence query Q, temporal grounding aims to determine the start and end timestamps (τ^s, τ^e) of specific video moment semantically corresponding to the sentence query. The video V is represented as $\boldsymbol{V} = \{\boldsymbol{v}_t\}_{t=1}^{T}$ frame-by-frame, and the query is represented as $\boldsymbol{Q} = \{\boldsymbol{q}_n\}_{n=1}^{N}$ word-by-word, where T and N are the number of frames and words, respectively.

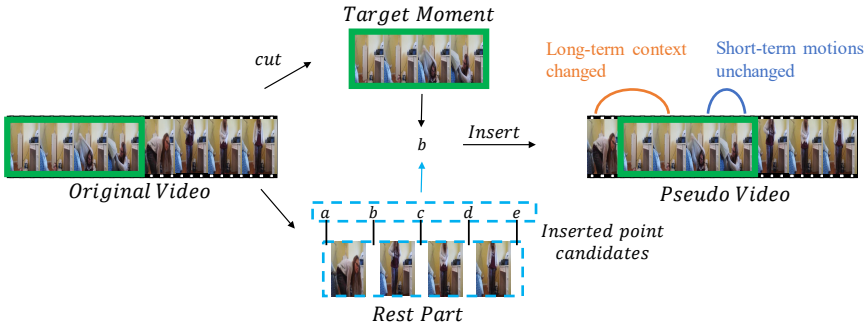


Fig. 2: An illustration of the generation of pseudo videos. The inserted point is randomly sampled from the candidates.

3.2 Input Construction

We first introduce how we shuffle videos to construct the input of our framework. For each video-query pair in the training set, we first cut out the target moment from the video and then insert the cut moment into a random temporal position of the rest of the video, as shown in Fig. 2. We name the shuffled videos pseudo videos. The pseudo videos have three characteristics: 1) temporal positions of target moments do not match the temporal biases of queries; 2) spatial contents and short-term temporal motions within target moments are consistent with original videos; 3) long-term temporal contexts around target moments are disturbed. Leveraging the three characteristics of pseudo videos, we design two auxiliary tasks to suppress the effect of temporal biases on the final reasoning and strengthen the grounding model's perception ability on visual contents.

Formally, for each video-query pair (V, Q) in the training set, we construct a triplet $(V, \bar{V}, Q)$ as the input. $\bar{\boldsymbol{V}} = \{\bar{\boldsymbol{v}}_t\}_{t=1}^{T}$ denotes the generated pseudo video and the corresponding timestamps of the target moment is $(\bar{\tau}^s, \bar{\tau}^e)$. Pseudo video $\bar{V}$ has the same length as original video V.

3.3 Framework Architecture

As shown in Fig. 3, our model consists of a grounding model, a cross-modal semantic matching module, and a temporal order discriminator. The grounding model predicts the locations of target moments. The latter two modules aim to address two auxiliary tasks, cross-modal matching and temporal order discrimination, respectively. The cross-modal semantic matching module predicts the relevance to the query for each video frame. And the predicted frame-level relevance scores will be used to gate the encoded video features in the span predictor. The temporal order discriminator predicts whether the input video is in correct order. The three modules share a video encoder so that the auxiliary tasks can promote the grounding model training.

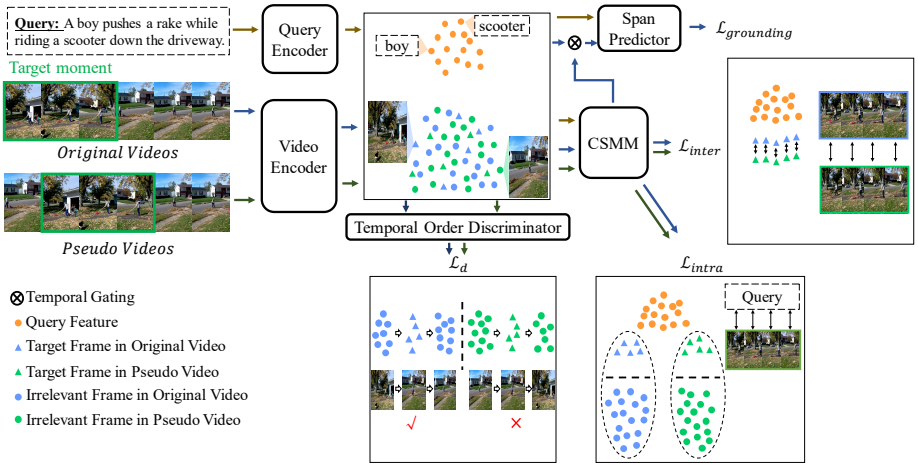


Fig. 3: An overview of our proposed framework with two auxiliary tasks. Video and query encoder encodes video and language modalities, respectively. The span predictor predicts the boundary scores for each frame. CSMM denotes the cross-modal semantic module, which predicts the relevance to the query for each video frame. We constrain the predict scores for both intra- and inter-video to deeply mine the visual-textual semantics relevance. Temporal order discriminator classifies whether the video moment sequence is in correct order.

3.4 Cross-Modal Matching Task

The cross-modal matching task aims to strengthen the visual-textual matching based on the content consistency between pseudo and original videos. We design two losses to constrain the predicted relevance scores for intra- and inter-video, respectively. For intra-video, the cross-modal semantic matching module should discriminate which frames are semantically related to the query for both pseudo and original videos. We implement this through two binary cross-entropy losses,

$$\mathcal{L}_{\text{BCE}}(\boldsymbol{V}) = -\sum_{\boldsymbol{v}_t}^{\boldsymbol{V}} p(\boldsymbol{v}_t)\log(c(\boldsymbol{v}_t)) + (1-p(\boldsymbol{v}_t))\log(1-c(\boldsymbol{v}_t)), \tag{1}$$

$$\mathcal{L}_{intra} = \frac{1}{2}(\mathcal{L}_{\text{BCE}}(\boldsymbol{V}) + \mathcal{L}_{\text{BCE}}(\bar{\boldsymbol{V}})), \tag{2}$$

where $p(\boldsymbol{v}_t)$ is set to 1 if frame $\boldsymbol{v}_t$ is within the target moments and 0 for otherwise, $c(\boldsymbol{v}_t)$ denotes the predicted cross-modal relevance scores for frame $\boldsymbol{v}_t$ with sigmoid activation. For the same query, target frames in pseudo and original videos may be at different temporal positions. Thus the model has to mine visual contents semantically matching the query. Since the long-term temporal context in videos may be incorrect, this loss encourages the model to reason cross-modal relevance more based on spatial contents and short-term temporal motions.

For inter-video, we constrain that the predicted relevance distributions within target moments should be consistent between pseudo and original videos. We use a Kullback-Leibler divergence to constrain the fine-grained consistency,

$$\mathcal{L}_{inter} = D_{\mathrm{KL}}(\boldsymbol{c} \,||\, \bar{\boldsymbol{c}}), \tag{3}$$

where $\boldsymbol{c}$ denotes the softmaxed relevance score vector of frames $\boldsymbol{v}_t$ from τ^s to τ^e timestep in video $\boldsymbol{V}$ and $\bar{\boldsymbol{c}}$ denotes the softmaxed relevance score vector of frames $\bar{\boldsymbol{v}}_t$ from $\bar{\tau}^s$ to $\bar{\tau}^e$ timestep in pseudo video $\bar{\boldsymbol{V}}$. This inter-video loss constrains the relative relevance differences within target moments unchanged even though the external temporal context is changed. Thus this loss further emphasizes the impact of spatial and short-term temporal motion features.

3.5 Temporal Order Discrimination Task

The cross-modal matching task emphasizes the impact of spatial and short-term temporal motion features in final prediction, but we wish the grounding model is capable of understanding the long-term temporal contexts, which is important for temporal grounding task [17,60,59,44]. However, some methods [17,60,45,12] learn the potential temporal position information during context capturing and thus suffer the temporal bias problem [37,51]. To this end, we introduce a temporal order discrimination task to guide explicitly the learning of long-term temporal contexts. This task aims to discriminate whether the input video is in correct temporal order. Unlike the existing temporal order tasks [10,49] that focus on the order of frames sampled from a short-span action, we design a task to focus on the long-term temporal context. Specifically, given a video-query pair, we divide the video into three parts: the target moment, the moment before the target moment, and the moment after the target moment, and ask whether the three moments are correctly ordered. We determine the supervision for this task based on the video type, i.e., we suppose that the orders of the original videos are correct and the ones of the pseudo videos are incorrect. This task is trained by a cross-entropy loss, which is denoted as $\mathcal{L}_d$.

$$L_d = -\sum_{c=1}^{C} y_{V,c} \log o_c(V) - \sum_{c=1}^{C} y_{\bar{V},c} \log o_c(\bar{V}) \tag{4}$$

where C denotes the video categories (original or pseudo), y is groundtruth label and $o_c(V)$ denotes the softmaxed prediction score of video V for category c.

3.6 Span-based Grounding Model

We apply a span-based grounding model [14] as our baseline model. It applies a typical span-based architecture, consisting of a query encoder, a video encoder, and a span predictor[1]. Query encoder models the sentence query with

[1] As the grounding model is not our key contribution, more details about the grounding network and inference stage are provided in our supplementary material.

multi-layered bidirectional LSTM with pre-trained language model (GloVe [38]) embeddings as input. The encoded word-level embeddings and sentence-level representation are denoted as $\boldsymbol{W} = \{\dot{\boldsymbol{w}}_n\}_{n=1}^{N}$ and $\boldsymbol{s}$, respectively. The video encoder is guided by the query to encode video features over time. Different from [14], we only use the word-level features $\boldsymbol{W}$ to guide the video encoding.

$$\boldsymbol{W}, \boldsymbol{s} = \text{QueryEncoder}(\boldsymbol{Q}), \quad \dot{\boldsymbol{V}} = \text{VideoEncoder}(\boldsymbol{V}, \boldsymbol{W}). \tag{5}$$

Each encoded frame feature $\dot{\boldsymbol{v}}$ is concatenated with the sentence representation $\boldsymbol{s}$ and gated by the cross-modal relevance score $c(\boldsymbol{v}_t)$ before feed forward to the span predictor. The span predictor predicts the boundary scores $S_{start}(t)$, $S_{end}(t)$ for each frame.

$$\begin{aligned} S_{start}(t) &= \text{StartPredictor}(c(\boldsymbol{v}_t)(\dot{\boldsymbol{v}}_t||\boldsymbol{s})), \\ S_{end}(t) &= \text{EndPredictor}(c(\boldsymbol{v}_t)(\dot{\boldsymbol{v}}_t||\boldsymbol{s})). \end{aligned} \tag{6}$$

The start and end scores are normalized with SoftMax to obtain $P_{start}(t)$, $P_{end}(t)$, and trained using negative log-likelihood loss:

$$\mathcal{L}_g = -\log(P_{start}(t^s)) - \log(P_{end}(t^e)), \tag{7}$$

where t^s, t^e are the ground-truth start and end frame indices for the original video, respectively. The ground-truth frame indices (t^s, t^e) are mapped from the ground-truth time values (τ^s, τ^e).

3.7 Cross-Modal Semantic Matching Module

The cross-modal semantic matching module predicts the relevance to the query for each video frame. The module is implemented by a multi-layered perceptron (MLP) with relu activation in hidden layers. We also use concatenation as the cross-model interaction method,

$$c(\boldsymbol{v}_t) = \boldsymbol{W}_2^c \text{relu}(\boldsymbol{W}_1^c(\dot{\boldsymbol{v}}_t||\boldsymbol{s}) + \boldsymbol{b}_1^c) + \boldsymbol{b}_2^c, \tag{8}$$

where $\boldsymbol{W}_1^c$, $\boldsymbol{W}_2^c$, $\boldsymbol{b}_1^c$ and $\boldsymbol{b}_2^c$ are the learnable parameters of MLP and shared across all time steps. We also apply a temporal gating on the encoded video features using the predicted relevance scores to highlight the impact of the matching results on the final reasoning, as shown in Eqs. (6).

3.8 Temporal Order Discriminator

Given a moment tuple ($M_1 = \{\dot{\boldsymbol{v}}_1, \cdots, \dot{\boldsymbol{v}}_{\tau^s-1}\}, M_2 = \{\dot{\boldsymbol{v}}_{\tau^s}, \cdots, \dot{\boldsymbol{v}}_{\tau^e}\}, M_2 = \{\dot{\boldsymbol{v}}_{\tau^e+1}, \cdots, \dot{\boldsymbol{v}}_T\}$), we first obtain the moment-level represenstations for these moments by average pooling the encoded frame features within the moments,

$$\boldsymbol{m}_1 = \text{pooling}(M_1), \ \boldsymbol{m}_2 = \text{pooling}(M_2), \ \boldsymbol{m}_3 = \text{pooling}(M_3). \tag{9}$$

We mainly focus on the contexts around the target moment (M_2) to reason the order correctness. We concatenate the paired moment representations and use two parallel fully-connected layers with shared parameters to obtain the context information $\boldsymbol{h}_1$, $\boldsymbol{h}_2$, respectively,

$$\boldsymbol{h}_1 = \text{relu}(\boldsymbol{W}_1^o(\boldsymbol{m}_1||\boldsymbol{m}_2 + \boldsymbol{b}_1^o)),\ \boldsymbol{h}_2 = \text{relu}(\boldsymbol{W}_1^o(\boldsymbol{m}_2||\boldsymbol{m}_3 + \boldsymbol{b}_1^o)). \tag{10}$$

Then we concatenate the context information with the target moment representation and predict the classification scores,

$$o(\boldsymbol{V}) = \boldsymbol{W}_2^o(\boldsymbol{m}_2||\boldsymbol{h}_1||\boldsymbol{h}_2) + \boldsymbol{b}_2^o, \tag{11}$$

where $\boldsymbol{W}_1^o$, $\boldsymbol{W}_2^o$, $\boldsymbol{b}_1^o$ and $\boldsymbol{b}_2^o$ are the learnable parameters of fully-connected layers and shared across videos. To prevent the learning of temporal biases, we reason the order correctness only based on the content relevance and do not introduce any global temporal position information. The position information contained in the encoded frame features is diluted by the average-pooling operation.

3.9 Training Objective

The final training loss is the weighted summarization of the loss of each module,

$$\mathcal{L} = \mathcal{L}_g + \lambda_1\mathcal{L}_{intra} + \lambda_2\mathcal{L}_{inter} + \lambda_3\mathcal{L}_d. \tag{12}$$

4 Experiments

4.1 Datasets

Charades-STA Charades-STA is built on Charades dataset [40] by [11]. The videos in this dataset are mainly about indoor activities. The average length of videos and annotated moments are 30s and 8s, respectively.

ActivityNet Captions ActivityNet Captions is built on ActivityNet [15], which is a large-scale dataset of human activities based on YouTube videos. The average length of videos and the annotated moments are 117s and 36s, respectively.

4.2 Dataset Splits

In the original splits of the two datasets, the training and test sets have similar temporal biases. Thus the methods that learn the temporal biases in the training set could also perform well on the test set [37,52]. To eliminate the impact of temporal bias problem on evaluation performance, we perform experiments on the re-divided splits[2] proposed by [52]. In the re-divided splits, each dataset is re-divided into four sets: training, validation(val), test-iid, and test-ood. The temporal locations of all samples in the training, val, and test-iid satisfy the

[2] `https://github.com/yytzsy/grounding_changing_distribution`

Table 1: The statistics of the number of videos and query-moment pairs in different datasets and splits.

Dataset	Original Splits			Re-divided Splits		
	Split	Videos	Pairs	Split	Videos	Pairs
Charades-STA	training	5,338	12,408	training	4,564	11,071
	test	1,334	3,720	val	333	859
				test-iid	333	823
				test-ood	1,442	3,375
ActivityNet Captions	training	10,009	37,421	training	10,984	51,415
	val	4,917	17,505	val	746	3,521
	test	4,885	17,031	test-iid	746	3,443
				test-ood	2,450	13,578

independent and identical distribution, and the samples in test-ood are out-of-distribution. Therefore, it is useless to exploit the temporal biases in training set to make predictions on test-ood set. The test-ood sets on both datasets have similar vocabulary distributions to the training set, which means that the difference of the temporal distribution between the training and test-ood sets is the main challenge of the re-divided splits. The sample statistics of the two splits are reported in Table 1.

4.3 Experimental Settings

Metrics Following the conventions, we adopt R@n, IoU=θ and mIoU as evaluation metrics. R@n, IoU=θ represents the percentage of testing samples having at least one result whose IoU with ground truth is larger than θ in top-n localized results. mIoU represents the average IoU over all testing samples. Following previous works [13,45,58], we use $n = 1$ and $\theta \in \{0.3, 0.5, 0.7\}$.

Implementation For natural language, we use 300d Glove [38] vectors as word embeddings. For the words not in the vocabulary of Glove, we generate their embeddings randomly. For video modality, we use 1024d I3D feature pre-trained on Kinetics dataset [2] or 500d C3D feature [41] pre-trained on Sports-1M dataset [20] as the initial frame features, and downsample the videos at a frame rate of 1 frame per second. For each training epoch, we will re-generate a pseudo video for each video-query pair. We train the model with a batch size of 32 for 30 epochs for all datasets using Adam optimizer with an initial learning rate of 0.001. We set λ_1, λ_2 and λ_3 in Eq. 12 to 1, 1, 1, respectively. More details are provided in supplementary material.

4.4 Comparison with State-of-the-Arts

We compare our methods with the recently state-of-the-art methods and our baseline, which only contains a span-based grounding model [14] and is only supervised by the grounding loss $\mathcal{L}_g$. We first analyze the competitors' performance on the re-divided splits. Then we follow [37] and perform a test to check whether the competitors suffer the temporal bias problem. We also show the performance comparison on the original splits.

Table 2: Comparison on Charades-CD split using I3D features.

Model	test-iid				test-ood			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
2D-TAN [60]	60.15	49.09	26.85	42.73	52.79	35.88	13.91	34.22
LG [34]	64.52	51.28	28.68	45.16	59.32	42.90	19.29	39.43
DRN [55]	53.22	42.04	23.32	28.21	45.87	31.11	15.17	23.05
VSLNet [58]	61.48	43.26	28.43	42.92	54.61	34.10	17.87	36.34
DCM [51]	67.27	55.81	37.30	48.74	60.89	45.47	22.70	40.99
Baseline	66.34	50.55	34.26	46.81	58.96	38.22	20.50	39.52
Ours	**70.72**	**57.59**	**37.79**	**50.93**	**64.95**	**46.67**	**27.08**	**44.30**

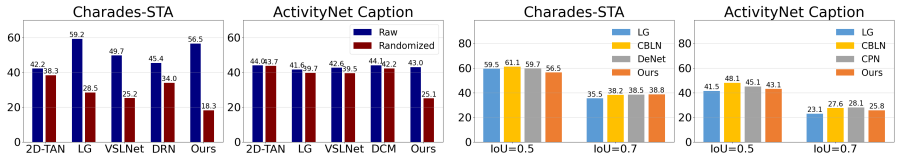
Table 3: Comparison on ActivityNet-CD split using I3D features.

Model	test-iid				test-ood			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
2D-TAN [60]	60.56	46.59	30.55	44.99	40.13	22.01	10.34	28.31
LG [34]	61.63	46.41	29.28	44.62	40.78	23.85	10.96	28.46
VSLNet [58]	62.71	47.81	29.07	46.33	38.30	20.03	10.29	28.18
DCM [51]	60.15	47.26	31.97	45.20	39.39	22.32	11.22	28.08
Baseline	60.56	44.58	27.42	44.28	38.78	21.39	10.86	28.41
Ours	**63.29**	**48.07**	**32.15**	**47.03**	**42.08**	**24.57**	**13.21**	**30.45**

Comparison on the Re-divided Splits. Table 2 and Table 3 summarize the results on the splits Charades-CD and ActivityNet-CD, respectively. For a fair comparison, all methods use the same pre-trained visual features. Best results are in **bold** and second-best underlined. Observed that all methods have a significant performance drop on the test-ood compared to the test-iid on both datasets. The change of the temporal distribution on the test-ood set challenges the model's generalization ability. Compared with the state-of-the-art methods, our method performs best over all evaluation metrics on both test sets and on both datasets. Particularly, our method outperforms all methods by clear margins on the test-ood splits, especially in the metrics IoU=0.7 and mIoU. It shows that our method has a stronger generalization ability against the temporal distributions. Besides, observed that our method improves the baseline on both test sets, but we achieve more relative improvements on test-ood set than test-iid set, e.g., 22.11% v.s. 13.93% in IoU=0.5 on Charades-CD and 14.87% v.s. 7.83% on ActivityNet-CD. It shows the effectiveness of our method on strengthening the model's generalization ability against the different temporal distributions.

Sanity Check on Visual Input. Suffering temporal bias problem, a model can ignore the visual input but perform well on the evaluation metrics on the original splits. Thus, same as [37], we perform a test on some state-of-the-art competitors[3] to show how much these models take input videos into account for prediction. Specifically, we divide input videos into short segments and randomly reorder them before evaluating the models. This randomization messes up the correspondence between input videos and ground truth temporal locations. If a model makes the prediction based on visual input, the performance should drop significantly by the randomization; otherwise, we can conclude that the

[3] We used the models with trained parameters provided by their authors if available.



(a) R@1 (IoU=0.5) scores for competitors when the raw input videos and randomized ones are fed into these models.

(b) Comparison with the state-of-the-arts on the original splits of Charades-STA and ActivityNet Captions.

Fig. 4: Sanity check on visual input (a) and comparison on the original splits (b)

Table 4: Comparison of different usage of pseudo videos on test-ood sets.

Methods	Charades-CD				ActivityNet-CD			
	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
Baseline	58.96	38.22	20.50	39.52	38.78	21.39	10.86	28.41
Baseline+DataAug	58.61	37.40	19.82	39.06	40.74	22.27	12.16	29.59
Baseline+CSMM+DataAug	59.22	38.45	21.41	39.84	28.42	13.36	6.13	20.29
Ours	64.95	46.67	27.08	44.30	42.08	24.57	13.21	30.45

model doesn't make use of the visual input during reasoning. Fig. 4a shows the results for the competitors and our method. On Charades-STA, all state-of-the-arts except 2D-TAN [60] and DRN [55] show significant performance drops for randomized videos. Among all competitors, our method performs the most significant drop, which demonstrates the effectiveness of our method for addressing temporal bias problem on this dataset. On ActivityNet Captions, all state-of-the-arts including DCM [51] achieve similar performance using randomized videos to the raw ones, which shows that these methods actually do not use the visual input and over-rely on the temporal biases. Our method performs a clear drop in this test on ActivityNet Captions, which validates that our method can make the model focus more on the visual content and mitigate the reliance on the temporal biases on this dataset.

Comparison on the Original Splits. We also compare our method with the state-of-the-arts [34,26,64,63] on the original splits of the two datasets. For a fair comparison, our method uses i3d features on Charades-STA and c3d on ActivityNet Captions. As shown in Fig. 4b, our method achieves competitive performance to the state-of-the-arts on both datasets, which shows that our method can address the temporal bias problem without loss in grounding accuracy.

4.5 Ablation Study

Pseudo videos. We study the effect of the pseudo videos. We first test applying the data augmentation strategy based on the baseline, i.e., treating the pseudo and original videos equally as training samples. Then we add the cross-modal matching module and use the frame-level relevance scores to temporally gating

Table 5: Ablation study of loss terms on test-ood sets.

Row	Loss Terms				Charades-CD				ActivityNet-CD			
	$\mathcal{L}_g$	$\mathcal{L}_{intra}$	$\mathcal{L}_{inter}$	$\mathcal{L}_d$	IoU=0.3	IoU=0.5	IoU=0.7	mIoU	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
1	✓				55.03	28.91	14.08	35.05	37.30	20.30	10.42	26.75
2	✓	✓			62.34	44.60	25.41	42.66	40.39	22.45	11.51	29.77
3	✓		✓		55.39	33.13	17.23	36.36	34.85	15.80	7.66	26.27
4	✓			✓	55.91	6.16	2.79	30.47	34.84	15.65	7.95	26.43
5	✓	✓		✓	62.96	45.36	26.73	43.40	41.49	23.12	12.55	30.04
6	✓	✓	✓		62.79	44.23	25.90	42.93	41.33	23.28	12.53	30.04
7	✓	✓	✓	✓	64.95	46.67	27.08	44.30	42.08	24.57	13.21	30.45

the encoded video features. The relevance scores are supervised by Eq.(1). As shown in Table 4, after applying data augmentation, there are slight improvements from the baseline on ActivityNet-CD while the performance is inferior to the baseline on Charades-CD. After adding CSMM, there are slight improvements on Charades-CD but a significant performance drop on ActivityNet-CD. On the contrary, our method improves the performance from the baseline with clear margins on both datasets. The comparisons validate the infeasibility of shuffled videos as augmented training samples and the superiority of our method.

Loss terms. We analyze the impact of each loss term and their combinations. Table 5 summarizes the results, and some reveal points are listed as follows.

(1) $\mathcal{L}_{intra}$ leads to the main performance boosts on test-ood sets (comparing Rows 1, 2 and 7). It validates our design of using the content consistency between shuffled and original videos to mitigate the reliance on temporal biases.

(2) Without $\mathcal{L}_{intra}$, the improvement of adding $\mathcal{L}_{inter}$ is limited (comparing Rows 1 and 3). It means that only constraining the relative relevance differences within target moments cannot highlight the target moment from the entire video. The combination of $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ can further improve the performance (comparing Rows 2 and 6), which validates the effectiveness of our design of contraining the cross-modal relevance scores between shuffled and original videos.

(3) Only adding $\mathcal{L}_d$ leads to performance drops on both datasets, especially on the high IoU (comparing Rows 1 and 4). The baseline over-relies on memorizing temporal biases but the temporal order discrimination task restricts the learning of temporal position information. So the baseline model degrades to predict long-span predictions[4]. After adding the cross-modal matching task, the supervision guides the model to focus on the short-term visual contents semantically matching queries and thus leads to performance boosts (Rows 5 and 7).

Visualization. We show a qualitative example on Charades-CD to show how models suffer the temporal bias problem and the effect of our method in Fig. 5. We first show the temporal distribution comparison of the word 'undress' between the training/test_ood sets and the prediction results of grounding models

[4] The length distribution of predictions can be found in our supplementary material.

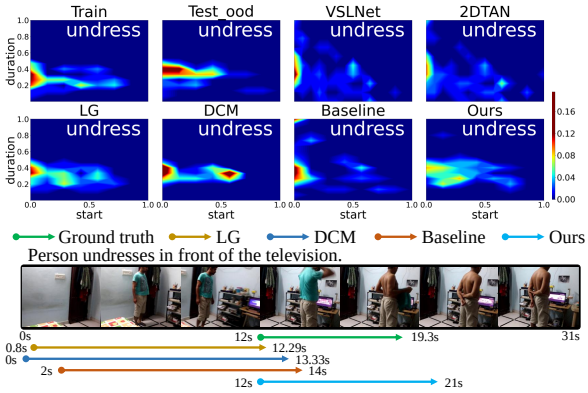


Fig. 5: Top is the temporal distribution comparison of the word 'undress' on Charades-CD. Color represents value of probability. Bottom is an example of grounding results for a query containing the word 'underss'.

on test_ood set. Observed that the competitor models have significant and similar biases to the training set while our model does not. Then we show a test samples of the word 'undress' and the grounding results of different methods. In this sample video, groundtruth target moment does not start at the beginning. But most models (LG [34], DCM [51], and our baseline) still make the predictions fitting the biases in training set. With the training of our framework, we can effectively mitigate the baseline's reliance on biases and turn the model's attention back to visual contents to make correct predictions. We provide more qualitative examples in our supplementary material.

5 Conclusion

This paper proposes a novel training framework for temporal grounding models to leverage shuffled videos to address the temporal bias problem. We propose two auxiliary tasks to suppress the effect of temporal biases and strengthen the model's perception ability on visual contents. Extensive experiments on Charades-STA and ActivityNet Captions demonstrate the effectiveness of our method on strengthening the generalization ability of the grounding model and mitigating the reliance on temporal biases.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants (62071067, 62001054, 62101064, 62171057), in part by the Ministry of Education and China Mobile Joint Fund (MCM20200202), China Postdoctoral Science Foundation under Grant 2022M710468, Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

1. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multi-modal transformer for video grounding. In: EMNLP. pp. 9810–9823 (2021)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR (2017)
3. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.: Temporally grounding natural sentence in video. In: EMNLP (2018)
4. Chen, L., Lu, C., Tang, S., Xiao, J., Zhang, D., Tan, C., Li, X.: Rethinking the bottom-up framework for query-based video localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10551–10558 (2020)
5. Chen, S., Jiang, W., Liu, W., Jiang, Y.: Learning modality interaction for temporal sentence localization and event captioning in videos. In: ECCV (2020)
6. Chen, S., Jiang, Y.: Semantic proposal for activity localization in videos via sentence query. In: AAAI (2019)
7. Chen, S., Jiang, Y.G.: Hierarchical visual-textual graph for temporal activity localization via language. In: ECCV (2020)
8. Chen, Y.W., Tsai, Y.H., Yang, M.H.: End-to-end multi-modal video temporal grounding. NIPS **34** (2021)
9. Choi, J., Sharma, G., Schulter, S., Huang, J.: Shuffle and attend: Video domain adaptation. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII. pp. 678–695 (2020)
10. Fernando, B., Bilen, H., Gavves, E., Gould, S.: Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3636–3645 (2017)
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: ICCV (2017)
12. Gao, J., Xu, C.: Fast video moment retrieval. In: ICCV. pp. 1523–1532 (2021)
13. Ghosh, S., Agarwal, A., Parekh, Z., Hauptmann, A.G.: Excl: Extractive clip localization using natural language descriptions. In: NAACL (2019)
14. Hao, J., Sun, H., Ren, P., Wang, J., Qi, Q., Liao, J.: Query-aware video encoder for video moment retrieval. Neurocomputing (2022)
15. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
16. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with natural language. In: ICCV (2017)
17. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with temporal language. In: EMNLP (2018)
18. HOU, Z., NGO, C.W., CHAN, W.: Conquer: Contextual query-aware ranking for video corpus moment retrieval.(2021). In: ACM MM. pp. 20–24 (2021)
19. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel. pp. 495–502 (2010)
20. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)
21. Lee, H., Huang, J., Singh, M., Yang, M.: Unsupervised representation learning by sorting sequences. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 667–676 (2017)

22. Lin, Z., Zhao, Z., Zhang, Z., Zhang, Z., Cai, D.: Moment retrieval via cross-modal interaction networks with query reconstruction. IEEE TIP (2020)
23. Liu, B., Yeung, S., Chou, E., Huang, D., Fei-Fei, L., Niebles, J.C.: Temporal modular networks for retrieving complex compositional activities in videos. In: ECCV (2018)
24. Liu, D., Qu, X., Di, X., Cheng, Y., Xu, Z., Zhou, P.: Memory-guided semantic learning network for temporal sentence grounding. arXiv preprint arXiv:2201.00454 (2022)
25. Liu, D., Qu, X., Dong, J., Zhou, P.: Adaptive proposal generation network for temporal sentence localization in videos. In: EMNLP. pp. 9292–9301 (2021)
26. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: CVPR. pp. 11235–11244 (2021)
27. Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: ACM MM (2020)
28. Liu, D., Qu, X., Zhou, P.: Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In: EMNLP. pp. 9302–9311 (2021)
29. Liu, D., Qu, X., Zhou, P., Liu, Y.: Exploring motion and appearance information for temporal sentence grounding. arXiv preprint arXiv:2201.00457 (2022)
30. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.: Attentive moment retrieval in videos. In: SIGIR (2018)
31. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.: Cross-modal moment localization in videos. In: ACM MM (2018)
32. Lu, C., Chen, L., Tan, C., Li, X., Xiao, J.: DEBUG: A dense bottom-up grounding approach for natural language video localization. In: EMNLP (2019)
33. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)
34. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: CVPR (June 2020)
35. Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: CVPR. pp. 2765–2775 (2021)
36. Ng, J.Y., Hausknecht, M.J., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR. pp. 4694–4702 (2015)
37. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J.: Uncovering hidden challenges in query-based video moment retrieval. In: BMVC (2020)
38. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
39. Rodriguez, C., Marrese-Taylor, E., Saleh, F.S., Li, H., Gould, S.: Proposal-free temporal moment localization of a natural-language query in video using guided attention. In: WACV (2020)
40. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016)
41. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
42. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)

43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
44. Wang, H., Zha, Z.J., Li, L., Liu, D., Luo, J.: Structured multi-level interaction network for video moment localization via language query. In: CVPR. pp. 7026–7035 (2021)
45. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: AAAI (2020)
46. Wang, L., Tong, Z., Ji, B., Wu, G.: TDN: temporal difference networks for efficient action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 1895–1904 (2021)
47. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 7794–7803 (2018)
48. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 2986–2994 (2021)
49. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 10334–10343 (2019)
50. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: AAAI (2019)
51. Yang, X., Feng, F., Ji, W., Wang, M., Chua, T.: Deconfounded video moment retrieval with causal intervention. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 1–10 (2021)
52. Yuan, Y., Lan, X., Chen, L., Liu, W., Wang, X., Zhu, W.: A closer look at temporal sentence grounding in videos: Datasets and metrics. CoRR **abs/2101.09028** (2021), `https://arxiv.org/abs/2101.09028`
53. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: NIPS (2019)
54. Yuan, Y., Mei, T., Zhu, W.: To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI (2019)
55. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: CVPR (June 2020)
56. Zhang, D., Dai, X., Wang, X., Wang, Y., Davis, L.S.: MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In: CVPR (2019)
57. Zhang, H., Sun, A., Jing, W., Zhen, L., Zhou, J.T., Goh, R.S.M.: Natural language video localization: A revisit in span-based question answering framework. IEEE transactions on pattern analysis and machine intelligence (2021)
58. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: ACL (2020)
59. Zhang, M., Yang, Y., Chen, X., Ji, Y., Xu, X., Li, J., Shen, H.T.: Multi-stage aggregated transformer network for temporal language localization in videos. In: CVPR. pp. 12669–12678 (2021)
60. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI (2020)

61. Zhang, Z., Zhao, Z., Zhang, Z., Lin, Z., Wang, Q., Hong, R.: Temporal textual localization in video via adversarial bi-directional interaction networks. IEEE TMM (2020)
62. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019. pp. 655–664 (2019)
63. Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: CVPR. pp. 4197–4206 (2021)
64. Zhou, H., Zhang, C., Luo, Y., Chen, Y., Hu, C.: Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In: CVPR. pp. 8445–8454 (2021)