

# Transform-Equivariant Consistency Learning for Temporal Sentence Grounding

DAIZONG LIU, Huazhong University of Science and Technology, China  
XIAOYE QU, Huazhong University of Science and Technology, China  
JIANFENG DONG, Zhejiang Gongshang University, China  
PAN ZHOU, Huazhong University of Science and Technology, China  
ZICHUAN XU, Dalian University of Technology, China  
HAOZHAO WANG, Huazhong University of Science and Technology, China  
XING DI, Protagolabs Inc., USA  
WEINING LU, Tsinghua University, China  
YU CHENG, Microsoft Research, USA

This paper addresses the temporal sentence grounding (TSG). Although existing methods have made decent achievements in this task, they not only severely rely on abundant video-query paired data for training, but also easily fail into the dataset distribution bias. To alleviate these limitations, we introduce a novel Equivariant Consistency Regulation Learning (ECRL) framework to learn more discriminative query-related frame-wise representations for each video, in a self-supervised manner. Our motivation comes from that the temporal boundary of the query-guided activity should be consistently predicted under various video-level transformations. Concretely, we first design a series of spatio-temporal augmentations on both foreground and background video segments to generate a set of synthetic video samples. In particular, we devise a self-refine module to enhance the completeness and smoothness of the augmented video. Then, we present a novel self-supervised consistency loss (SSCL) applied on the original and augmented videos to capture their invariant query-related semantic by minimizing the KL-divergence between the sequence similarity of two videos and a prior Gaussian distribution of timestamp distance. At last, a shared grounding head is introduced to predict the transform-equivariant query-guided segment boundaries for both the original and augmented videos. Extensive experiments on three challenging datasets (ActivityNet, TACoS, and Charades-STA) demonstrate both effectiveness and efficiency of our proposed ECRL framework.

CCS Concepts: • **Information systems** → **Multimedia and multimodal retrieval**; *Video search*.

Additional Key Words and Phrases: Temporal sentence grounding, transformation, equivariant, consistency learning

## ACM Reference Format:

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Zichuan Xu, Haozhao Wang, Xing Di, Weining Lu, and Yu Cheng. 2023. Transform-Equivariant Consistency Learning for Temporal Sentence Grounding. *J. ACM* 37, 4, Article 111 (August 2023), 17 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

---

Authors' addresses: Daizong Liu, Huazhong University of Science and Technology, China; Xiaoye Qu, Huazhong University of Science and Technology, China; Jianfeng Dong, Zhejiang Gongshang University, China; Pan Zhou, Huazhong University of Science and Technology, China; Zichuan Xu, Dalian University of Technology, China; Haozhao Wang, Huazhong University of Science and Technology, China; Xing Di, Protagolabs Inc., USA; Weining Lu, Tsinghua University, China; Yu Cheng, Microsoft Research, USA.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2023 Association for Computing Machinery.

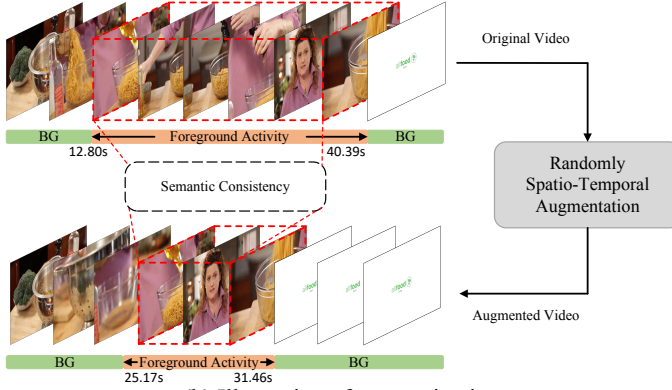
0004-5411/2023/8-ART111 \$15.00

<https://doi.org/XXXXXXXX.XXXXXXX>

**Query:** The woman adds and mixes in oil and spices while talking to the camera.



(a) An example of the temporal sentence grounding.



(b) Illustration of our motivation.

Fig. 1. (a) An illustrative example of TSG. (b) Illustration of our motivation. “BG” means the query-irrelevant background. Here, we learn to predict the transform-equivariant temporal boundaries of the query-related segments in both original and augmented videos.

## 1 INTRODUCTION

Temporal sentence grounding (TSG) is an important yet challenging task in video understanding [41, 43, 53–55, 89], which has drawn increasing attention over the last few years due to its vast potential applications in video captioning [7, 12, 29], video summarization [10, 77, 94], video-text retrieval [13, 14, 91], and video question answering [21, 35, 71], etc. As shown in Figure 1 (a), this task aims to ground the most relevant video segment according to a given sentence query. It is substantially more challenging as it needs to not only model the complex multi-modal interactions among video and query features, but also capture complicated context information for predicting the accurate query-guided segment boundaries.

Most previous works either follow a proposal-based framework [1, 5, 32, 34, 46, 47, 49–51, 96, 98, 99, 105, 107, 108] that first generates multiple segment proposals and then selects the most query-matched one, or follow a proposal-free framework [6, 63, 97, 102] that directly regresses the start and end timestamps of the segment with the multi-modal representations. Although they have achieved significant performance, these methods are data-hungry and require a large amount of annotated data for training. Moreover, recent studies [66, 95, 103] point out that existing works rely on exploiting the statistical regularities of annotation distribution for segment prediction, thus easily sticking into the substantial distribution bias existed in benchmark datasets. Therefore, how to synthesize positive video-query data pairs without human labors while mitigating the data distribution bias during the model training is an emerging issue for TSG task.

In this paper, we propose a novel Equivariant Consistency Regulation Learning (ECRL) framework to address the above issues in a self-supervised manner. As shown in Figure 1 (b), given an untrimmed video that contains a specific segment semantically corresponding to a query, the target segment boundaries would drastically change when applying transformations to the raw video (e.g., random upsampling or downsampling on two background sub-videos and one foreground

segment, respectively). This operation does not disrupt the main contents of the video, since the query-aware semantic of target segment is invariant to these transforms. Therefore, synthesizing a set of such augmented videos is more representative than the truncated ones [103], and can be utilized to assist the model training. Moreover, the stability and generalization ability of the model are crucial for precisely de-limiting the segment boundaries due to the existence of the data distribution bias. To make our ECRL be strongly equivariant with respect to numerous transforms, we further capture the consistent query-related semantics between the original and augmented videos for better discriminating foreground-background frame-wise representations.

To this end, we first introduce a spatio-temporal transformation strategy to apply various video-level augmentations on each video to synthesize new video samples. Considering that general transformations may destroy the continuity of the adjacent frames, we further propose a self-refine module to enhance the completeness and smoothness of the augmented video. Then, we present a novel self-supervised consistency loss (SSCL), which optimizes the frame-wise representations by minimizing the KL-divergence between the sequence similarity of original/augmented videos and a prior Gaussian distribution for capturing the consistent query-related semantics between two videos. At last, a shared grounding head is utilized to predict the transform-equivariant query-guided segment boundaries. In this manner, our ECRL not only can be well-trained with the enriched data samples but also is robust to the data distribution bias.

The main contributions of this work are three-fold:

- To our best knowledge, this paper represents the first attempt to explore transform-equivariance for TSG task. Specifically, we propose the novel Equivariant Consistency Regulation Learning (ECRL) framework, to capture the consistency knowledge between the original video and its spatio-temporal augmented variant.
- We propose a self-refine module to smooth the discrete adjacent frames of augmented video. Besides, we propose a self-supervised consistency loss (SSCL) to utilize KL-divergence with a prior Gaussian distribution to discriminate frame-wise representation for learning the invariant query-related visual semantic.
- Comprehensive evaluations are conducted on three challenging TSG benchmarks: ActivityNet, TACoS, and Charades-STA. Our method re-calibrates the state-of-the-art performance by large margins.

## 2 RELATED WORK

**Temporal action localization.** Temporal action localization is a task that involves classifying action instances by predicting their start and end timestamps along with their respective action category labels. This is a single-modal task that has been extensively studied in the literature [69, 78, 100]. Researchers have proposed two main categories of methods for temporal action localization, namely one-stage and two-stage methods [37, 86]. One-stage methods predict both the action boundaries and labels simultaneously. For instance, Xu *et al.* [88] used a graph convolutional network to perform one-stage action localization. On the other hand, two-stage methods first generate action proposals and then refine and classify confident proposals. Usually, the confident proposals are generated using the anchor mechanism [86, 90]. However, there are other methods for generating proposals, such as sliding window [75], temporal actionness grouping [109], and combining confident starting and ending frames [36].

**Temporal sentence grounding.** Temporal sentence grounding (TSG) is a multimedia task that aims to semantically link a given sentence query with a specific video segment by identifying its temporal boundary. This task was introduced by [20] and [1]. TSG is considerably more challenging than temporal action localization, as it requires capturing both visual and textual information and

modeling the complex multi-modal interactions between them to accurately identify the target activity. Unlike temporal action localization, TSG involves identifying the semantic meaning of a sentence query and mapping it to a specific video segment. This requires understanding the context and meaning of the query and interpreting it in relation to the visual content of the video. Additionally, TSG needs to consider the complex interactions between the visual and textual modalities to accurately model the target activity. Various TSG algorithms [3, 5, 16–18, 22, 25, 27, 28, 38, 40, 42, 45, 48, 52, 56, 57, 60, 81, 84, 85, 96, 101, 108, 110, 111] have been proposed within the proposal-based framework, which first generates multiple segment proposals, and then ranks them according to the similarity between proposals and the query to select the best matching one. Traditional methods for temporal sentence grounding, such as [58] and [20], use video segment proposals to localize the target segment. These methods first sample candidate segments from a video and then integrate the query with segment representations using a matrix operation. However, these methods lack a comprehensive structure for effectively modeling multi-modal feature interactions. To address this limitation and more effectively mine cross-modal interactions, recent works such as [87], [8], [22], and [106] have proposed integrating the sentence representation with each video segment individually and then evaluating their matching relationships. By incorporating more fine-grained features from both the visual and textual modalities, these methods can better capture the complex interactions between them and achieve improved performance on TSG tasks. Although these methods achieve good performances, they severely rely on the quality of the proposals and are time-consuming. Without using proposals, recent works [6, 39, 62–64, 73, 97, 102] directly regress the temporal locations of the target segment. They do not rely on the segment proposals and directly select the starting and ending frames by leveraging cross-modal interactions between video and query. Specifically, they either regress the start/end timestamps based on the entire video representation or predict at each frame to determine whether this frame is a start or end boundary. However, recent studies [66, 95, 103] point out that both types of above works are limited by the issue of distribution bias in TSG datasets and models. In this paper, we propose a novel framework to alleviate the data bias in a self-supervised learning manner.

**Self-supervised learning.** Self supervised learning (SSL) has become an increasingly popular research area in recent years [11, 15, 23, 65, 104]. In the context of videos, SSL methods have focused on tasks such as inferring the future [26], discriminating shuffled frames [61], and predicting speed [2]. Some recent works [19, 31, 70, 93] have also used contrastive loss for video representation learning, where different frames in a video or different frames in other videos are treated as negative samples. Different from these methods, our goal is fine-grained temporal understanding of videos and we treat a long sequence of frames as input data. Moreover, since the neighboring frames in video have high semantic similarities, directly regarding these frames as negatives like above works may hurt the learning. To avoid this issue, we learn the consistency knowledge by minimizing the KL-divergence with a prior Gaussian distribution.

### 3 METHOD

#### 3.1 Problem Definition and Overview

Given an untrimmed video  $\mathcal{V}$  and a sentence query  $\mathcal{Q}$ , we represent the video as  $\mathcal{V} = \{v_t\}_{t=1}^T$  frame<sup>1</sup>-by-frame, where  $v_t$  is the  $t$ -th frame and  $T$  is the number of total frames. Similarly, the query with  $N$  words is denoted as  $\mathcal{Q} = \{q_n\}_{n=1}^N$  word-by-word. The TSG task aims to localize the start and end timestamps  $(\tau_s, \tau_e)$  of a specific segment in video  $\mathcal{V}$ , which refers to the corresponding

<sup>1</sup>In this paper, the frame is a general concept for an actual video frame or a video clip which consists of a few consecutive frames.

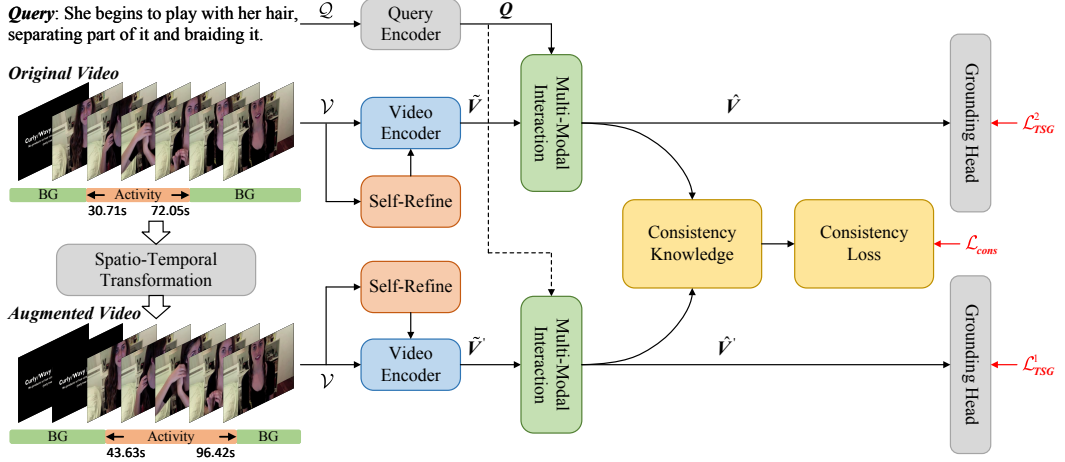


Fig. 2. Overall pipeline of the proposed ECRL. Given a pair of video and query input, we first apply spatio-temporal transformation on the original video to generate its augmented variant. Then, we encode both two videos with separate self-refine modules to enhance their completeness and interact them with the query. After that, we develop a self-supervised consistency learning module to discriminate the invariant query-relevant and -irrelevant frame-wise representations between two videos. At last, a shared grounding head is utilized to predict the transform-equivariant query-guided segment boundaries on them.

semantic of query  $Q$ . However, previous works not only rely on large amount of video-query pairs for training, but also tend to easily fit the data distribution bias during the model learning.

Therefore, we propose a novel Equivariant Consistency Regulation Learning (ECRL) framework to alleviate the above issues in a self-supervised manner. As shown in Figure 2, we first introduce a spatio-temporal transformation module to apply various video-level augmentations on each video to synthesize new video samples for assisting the model training. Considering the adjacent frames in augmented video tend to be discrete and incomplete, we further devise a self-refine module to enhance their smoothness. Note that, the new video samples not only can enrich the training data, but also can serve as contrastive samples for improving the model generalization ability. Therefore, we then present a novel self-supervised consistency loss (SSCL) to discriminate the frame-wise representations between both original and augmented videos for capturing their consistent query-related semantics. At last, a shared grounding head is utilized to predict the transform-equivariant query-guided segment boundaries on both original and augmented videos. We illustrate the details of each component in the following.

### 3.2 Spatio-Temporal Transformation

We first introduce the detailed spatio-temporal transformation step of our method to construct the augmented samples for better assisting the model learning. This data augmentation process is crucial to avoid trivial solutions in self-supervised learning [9]. Different from prior methods designed for image data which only require spatial augmentations, we introduce a series of spatio-temporal data augmentations to further increase the variety of videos.

**Temporal transformation.** For temporal data augmentation, since each video contains one specific segment corresponding to the sentence query, we first split each video into three parts (sub-video  $\mathcal{V}_{left}$  before the target segment, sub-video  $\mathcal{V}_{seg}$  of the target segment, and sub-video  $\mathcal{V}_{right}$  after the target segment) and then perform temporal transformations on them, respectively.

In detail, similar to the resize operation in image processing, we randomly perform up-sampling or down-sampling on each sub-video with different sampling ratio  $r_{left}, r_{seq}, r_{right} \in [1 - \alpha, \alpha]$  along the time dimension uniformly for changing their lengths. Particularly, as for the sub-video with 0 frame, empty frames are padded on it before applying the sampling operation. At last, we compose the three augmented sub-videos into a joint long video  $\mathcal{V}'$ , and uniformly *sample* it to the same length  $T$  as  $\mathcal{V}$  to generate the final augmented video.

**Spatial transformation.** For spatial data augmentation, we directly apply several spatial data augmentations, including random crop and resize, random color distortions, and random Gaussian blur, on video  $\mathcal{V}'$ .

### 3.3 Multi-Model Encoding and Interaction

**Video encoding.** For the original video  $\mathcal{V}$  and its augmented sample  $\mathcal{V}'$ , we first extract their frame-wise features by a pre-trained C3D network [79] as  $V = \{v_t\}_{t=1}^T, V' = \{v'_t\}_{t=1}^T \in \mathbb{R}^{T \times D}$ , where  $D$  is the feature dimension. Considering the sampled video sequence  $V'$  tends to be discrete and incomplete, we then introduce a **self-refine module** to utilize both *temporal* and *semantic* context information to smooth the consecutive frames. Specifically, we first construct a fully connected graph over  $V'$  where each node is a single frame. Let  $E \in \mathbb{R}^{T \times T}$  be the adjacency matrix of graph,  $E_{i,j}$  is the edge weight between node  $i$  and  $j$ . Intuitively, temporally neighboring frames are more likely to have correlated content. Therefore, we define the *temporal* adjacency weight as follows:

$$E_{i,j}^{tem} = e^{-\frac{|i-j|^2}{2\sigma^2}}, \quad (1)$$

where  $\sigma$  is empirically set as 5 in all experiments. For the *semantic* similarity of frames  $i$  and  $j$ , we directly evaluate it by measuring their cosine similarity as follows:

$$E_{i,j}^{sem} = \cos(v'_i, v'_j) = \frac{v'_i(v'_j)^\top}{\|v'_i\|_2 \|v'_j\|_2}. \quad (2)$$

The final inter-node similarity  $E_{i,j}$  is calculated by element-wise multiplication as  $E_{i,j} = E_{i,j}^{tem} \cdot E_{i,j}^{sem}$ . After that, we iteratively update and refine  $V'$  as follows:

$$\tilde{v}'_i = \sum_{j=1}^T E_{i,j} v'_j, \quad \tilde{v}'_i \in \tilde{V}' \quad (3)$$

where we empirically find that an overall iteration of 3 times will strike a good balance of accuracy and computational complexity. We can also enrich the self-contexts of  $V$  in the same manner and denote its final feature as  $\tilde{V}$ . At last, we employ a self-attention [80] layer and a BiLSTM [74] to capture the long-range dependencies within each video.

**Query encoding.** For the sentence query  $Q$ , we first generate the word-level features by using the Glove embedding [67], and then also employ a self-attention layer and a BiLSTM layer to further encode the query features as  $Q = \{q_n\}_{n=1}^N \in \mathbb{R}^{N \times D}$ .

**Multi-modal interaction.** After obtaining the encoded features  $\tilde{V}', \tilde{V}, Q$ , we utilize a co-attention mechanism [59] to capture the cross-modal interactions between video and query features. Specifically, for pair  $(\tilde{V}', Q)$ , we first calculate the similarity scores between  $\tilde{V}'$  and  $Q$  as:

$$S = \tilde{V}'(QW_S)^\top \in \mathbb{R}^{T \times N}, \quad (4)$$

where  $W_S \in \mathbb{R}^{D \times D}$  projects the query features into the same latent space as the video. Then, we compute two attention weights as:

$$A = S_r(QW_S) \in \mathbb{R}^{T \times D}, \quad B = S_r S_c^\top \tilde{V}' \in \mathbb{R}^{T \times D}, \quad (5)$$



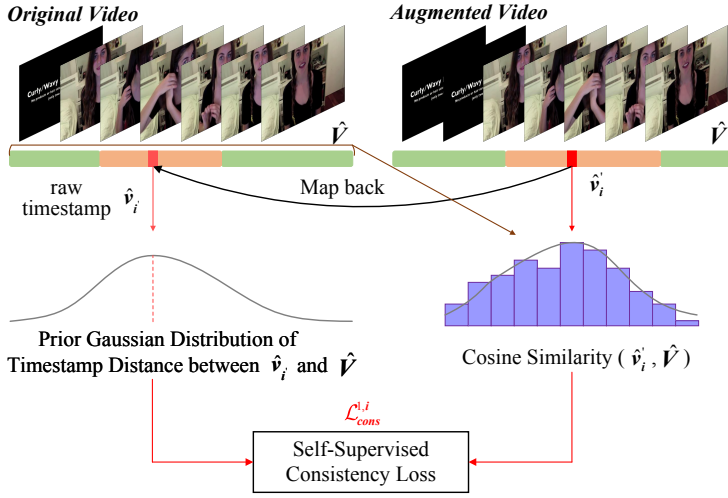


Fig. 3. Illustration of the proposed consistency loss. For  $\hat{v}'_i$  in  $\hat{V}'$ , we first compute a prior Gaussian distribution of timestamp distance between the raw timestamp  $i'$  of  $\hat{v}'_i$  and all timestamps in  $\hat{V}$  conditioned on the sub-video correspondence. Then the semantic similarity distribution between  $\hat{v}'_i$  and  $\hat{V}$  is calculated, and we minimize the KL-divergence of these two distributions for discriminating the frame-wise representations.

where  $S_r$  and  $S_c$  are the row- and column-wise softmax results of  $S$ , respectively. We compose the final query-guided video representation by learning its sequential features as:

$$\hat{V} = BiLSTM([\tilde{V}'; A; \tilde{V}' \odot A; \tilde{V}' \odot B]) \in \mathbb{R}^{T \times D}, \quad (6)$$

where  $\hat{V}' = \{\hat{v}'_t\}_{t=1}^T$ ,  $BiLSTM(\cdot)$  denotes the BiLSTM layers,  $[\cdot]$  is the concatenate operation, and  $\odot$  is the element-wise multiplication. In the same way, we can generate another query-guided video features  $\tilde{V}$  from the pair  $(\tilde{V}, Q)$ .

### 3.4 Transform-Equivariant Consistency Learning

Human visual perception shows good consistency for query-based segment localization when they watch the video at different playback rates. For example, when we watch a video that contains a specific activity, the corresponding semantic of the video segment will not change, yet the duration and temporal boundary of the segment will change as the playback rate varies. State differently, the query-related activity of the video is invariant to different playback rates, while its temporal boundary is equivariant. Therefore, to make our model have the equivariant property between the augmented video and its original one, we propose transform-equivariant consistency learning to maximize their agreements.

**How to learn the consistency knowledge?** For each video, the consistency knowledge denotes that the feature of the original frame in  $\mathcal{V}_{left}, \mathcal{V}_{seg}, \mathcal{V}_{right}$  should be semantically-invariant to the feature of the augmented frame in the same  $\mathcal{V}_{left}, \mathcal{V}_{seg}, \mathcal{V}_{right}$ , respectively. In this way, the model is able to discriminate the semantics of  $\mathcal{V}_{left}, \mathcal{V}_{seg}, \mathcal{V}_{right}$  with different playback rate, thus predicting the equivariant segment in the augmented video. To discriminate and learn the invariant frame-wise representations between two videos, a general idea is to take each corresponding frame as reference frame and take the other frames as negative ones. However, videos provide abundant sequential information, and the neighboring frames around the reference frame are highly correlated. Thus, directly regarding these frames (especially the segment boundaries) as negatives may hurt the

representation learning. To alleviate this issue, we present a novel self-supervised consistency loss (SSCL), which optimizes the frame-wise features by minimizing the Kullback–Leibler (KL) divergence [24] between the sequence similarity of two videos and a prior Gaussian distribution [68], to capture the consistency knowledge. As shown in Figure 3, to discriminate a single frame in  $\widehat{\mathbf{V}}'$  with the entire video  $\widehat{\mathbf{V}}$ , we first compute a prior Gaussian distribution of their timestamp distance in the original video, and then calculate the semantic similarity distribution between each single augmented frame and the entire original video. At last, we minimize the KL-divergence of the similarity distribution and the Gaussian distribution in the feature space.

**Formulation of the consistency loss.** Specifically, given the  $i$ -th augmented frame  $\widehat{v}_i'$  in  $\widehat{\mathbf{V}}'$ , we first find its raw video timestamp  $i'$  in original video  $\mathcal{V}$ . Since video  $\widehat{\mathbf{V}}$  is extracted from the original video, its timestamps  $\{1, 2, \dots, T\}$  are already the raw ones. Due to the fact that temporally adjacent frames are more highly correlated than those far away ones, we assume the similarity between  $\widehat{v}_i'$  and  $\widehat{\mathbf{V}}$  should follow a prior Gaussian distribution of timestamp distance between  $i'$  and  $\{1, 2, \dots, T\}$ .

Let  $G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$  denotes the Gaussian function, we use the KL-divergence to formulate the loss of  $i$ -th augmented frame in  $\widehat{\mathbf{V}}'$  as follows:

$$\mathcal{L}_{cons}^{1,i} = - \sum_j^T w_{ij} \log \frac{e^{\cos(\widehat{v}_i', \widehat{v}_j)}}{\sum_{t=1}^T e^{\cos(\widehat{v}_i', \widehat{v}_t)}}, \quad (7)$$

$$w_{ij} = \frac{G(i' - j)}{\sum_{t=1}^T G(i' - t)}, \quad (8)$$

where  $w_{ij}$  is the normalized Gaussian weight. *Note that*, for each frame in a sub-video, we apply an additional down-weight 0.5 to the Gaussian value of the frames in other sub-videos for better distinguishing. Similarly, we can calculate the loss of  $i$ -th original frame  $\mathcal{L}_{cons}^{2,i}$  for  $\widehat{\mathbf{V}}$ . Therefore, the overall SSCL loss function is formulated as:

$$\mathcal{L}_{cons} = \frac{1}{T} \sum_{i=1}^T (\mathcal{L}_{cons}^{1,i} + \mathcal{L}_{cons}^{2,i}). \quad (9)$$

### 3.5 Grounding Heads

To predict the target segments with the features  $\widehat{\mathbf{V}}'$ ,  $\widehat{\mathbf{V}}$  for both augmented and original videos, we employ the efficient proposal-free prediction head to regress the start and end timestamps of the segment. Specifically, for video  $\widehat{\mathbf{V}}'$ , we utilize two separate LSTM layers to successively predict the start and end scores on each video frame as:

$$\mathbf{h}_t^s = LSTM_{start}(\widehat{v}_t', \mathbf{h}_{t-1}^s), C_t^s = [\widehat{v}_t'; \mathbf{h}_t^s] \mathbf{W}_s + \mathbf{b}_s, \quad (10)$$

$$\mathbf{h}_t^e = LSTM_{end}(\widehat{v}_t', \mathbf{h}_{t-1}^e), C_t^e = [\widehat{v}_t'; \mathbf{h}_t^e] \mathbf{W}_e + \mathbf{b}_e, \quad (11)$$

where  $\mathbf{h}$  is the hidden state of LSTM layer,  $C_t^s, C_t^e$  denote the scores of start and end boundaries at  $t$ -th frame. We utilize the cross-entropy loss function  $\mathcal{L}_{ce}$  to supervise the grounding on the augmented video as:

$$\mathcal{L}_{TSG}^1 = \frac{1}{2T} \sum_{t=1}^T [\mathcal{L}_{ce}(C_t^s, \widehat{C}_t^s) + \mathcal{L}_{ce}(C_t^e, \widehat{C}_t^e)], \quad (12)$$

where  $\widehat{C}_t^s, \widehat{C}_t^e$  are the ground-truth labels. Similarly, we can formulate the loss function  $\mathcal{L}_{TSG}^2$  on the original video  $\widehat{\mathbf{V}}$ .



**Training.** During the training, we jointly optimize two grounding losses of two videos and the consistency loss as:

$$\mathcal{L}_{overall} = \mathcal{L}_{TSG}^1 + \mathcal{L}_{TSG}^2 + \lambda \mathcal{L}_{cons}, \quad (13)$$

where  $\lambda$  is the balanced weight.

**Testing.** During the inference, we directly construct the top-n segments by considering the summed scores of the selected start and end boundary timestamps for each video.

## 4 EXPERIMENTS

### 4.1 Dataset and Evaluation

**ActivityNet.** ActivityNet [30] contains 20000 untrimmed videos with 100000 descriptions from YouTube. Following public split, we use 37417, 17505, and 17031 sentence-video pairs for training, validation, testing.

**TACoS.** TACoS [72] is widely used on TSG task and contain 127 videos. We use the same split as [20], which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing.

**Charades-STA.** Charades-STA is built on [76], which focuses on indoor activities. As Charades dataset only provides video-level paragraph description, the temporal annotations of Charades-STA are generated in a semi-automatic way. In total, are 12408 and 3720 moment-query pairs in the training and testing sets.

**Evaluation.** Following previous works [20, 92], we adopt “R@n, IoU=m” as our evaluation metric, which is defined as the percentage of at least one of top-n selected moments having IoU larger than m.

### 4.2 Implementation Details

We implement our model in PyTorch. To extract video features, we utilize pre-trained C3D [79] to encode each video frames on ActivityNet, TACoS, and utilize pre-trained I3D [4] on Charades-STA. Since some videos are overlong, we pre-set the length  $T$  of video feature sequences to 200 for ActivityNet and TACoS datasets, 64 for Charades-STA dataset. As for sentence encoding, we set the length of word feature sequences to 20, and utilize Glove embedding [67] to embed each word to 300 dimension features. The hidden state dimension of BiLSTM networks is set to 512. The dimension  $D$  is set to 1024, and the weight  $\lambda$  is set to 5.0. During the video spatio-temporal transformation, we set the parameter  $\alpha = 0.8$ . During the training, we use an Adam optimizer with the leaning rate of 0.0001. The model is trained for 100 epochs to guarantee its convergence with a batch size of 64 (128 samples). All the experiments are implemented on a single NVIDIA TITAN XP GPU.

### 4.3 Comparison with State-of-the-Arts

**Compared methods.** We compare the proposed ECRL with state-of-the-art TSG methods on three datasets: TGN [5], CBP [81], SCDM [96], BpNet [83], CMIN [108], 2DTAN [105], DRN [98], CBLN [47], MMN [82], and MGSL [44], LGI [63], VSLNet [102], IVG-DCL [64].

**Quantitative comparison.** As shown in Table 1, we compare our proposed ECRL model with the existing TSG methods on three datasets, where our ECRL outperforms all the existing methods across different criteria by a large margin. Specifically, on ActivityNet dataset, compared to the previous best proposal-based method MGSL, we do not rely on large numbers of pre-defined proposals and outperform it by 2.37%, 1.56%, 2.17%, 1.64% in all metrics, respectively. Compared to the previous best bottom-up method IVG-DCL, our ECRL brings significant improvement of 10.40% and 5.88% in the R@1, IoU=0.5 and R@1, IoU=0.7 metrics. On TACoS dataset, the cooking activities take place in the same kitchen scene with some slightly varied cooking objects, thus it is hard to localize such fine-grained activities. Compared to the top ranked method MGSL, our model still

Table 1. Performance compared with the state-of-the-arts TSG methods on ActivityNet, TACoS, and Charades-STA datasets.

Method	ActivityNet				TACoS				Charades-STA			
	R@1,	R@1,	R@5,	R@5,	R@1,	R@1,	R@5,	R@5,	R@1,	R@1,	R@5,	R@5,
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
TGN	28.47	-	43.33	-	21.77	18.90	39.06	31.02	-	-	-	-
CBP	35.76	17.80	65.89	46.20	27.31	24.79	43.64	37.40	36.80	18.87	70.94	50.19
SCDM	36.75	19.86	64.99	41.53	26.11	21.17	40.16	32.18	54.44	33.43	74.43	58.08
LGI	41.51	23.07	-	-	-	-	-	-	59.46	35.48	-	-
BPNet	42.07	24.69	-	-	25.96	20.96	-	-	50.75	31.64	-	-
VSLNet	43.22	26.16	-	-	29.61	24.27	-	-	54.19	35.22	-	-
CMIN	43.40	23.88	67.95	50.73	24.64	18.05	38.46	27.02	-	-	-	-
IVG-DCL	43.84	27.10	-	-	38.84	29.07	-	-	50.24	32.88	-	-
2DTAN	44.51	26.54	77.13	61.96	37.29	25.32	57.81	45.04	39.81	23.25	79.33	51.15
DRN	45.45	24.36	77.97	50.30	-	23.17	-	33.36	53.09	31.75	89.06	60.05
CBLN	48.12	27.60	79.32	63.41	38.98	27.65	59.96	46.24	61.13	38.22	90.33	61.69
MMN	48.59	29.26	79.50	64.76	39.24	26.17	62.03	47.39	47.31	27.28	83.74	58.41
MGSL	51.87	31.42	82.60	66.71	42.54	32.27	63.39	50.13	63.98	41.03	93.21	63.85
<b>ECRL</b>	<b>54.24</b>	<b>32.98</b>	<b>84.77</b>	<b>68.35</b>	<b>45.20</b>	<b>34.43</b>	<b>65.74</b>	<b>51.86</b>	<b>65.37</b>	<b>42.69</b>	<b>94.52</b>	<b>65.18</b>

Table 2. Performance comparison on ActivityNet-CD and Charades-CD datasets [33].

Method	ActivityNet-CD		Charades-CD	
	test-IID	test-OOD	test-IID	test-OOD
Zhang <i>et al.</i> [103]	28.11	14.67	38.87	32.70
Lan <i>et al.</i> [33]	31.44	11.66	34.71	22.70
<b>ECRL</b>	<b>34.37</b>	<b>19.95</b>	<b>41.59</b>	<b>34.98</b>

Table 3. Efficiency comparison in terms of second per video (SPV) and parameters (Para.), where our method ECRL is much efficient with relatively lower model size.

	ACRN	CTRL	TGN	2DTAN	MGSL	VLSNet	<b>ECRL</b>
SPV ↓	4.31	2.23	0.92	0.57	0.10	0.07	<b>0.06</b>
Para. ↓	128	<b>22</b>	166	232	203	48	54

achieves the best results on the strict metrics R@1, IoU=0.5 and R@5, IoU=0.5 by boosting 2.16% and 1.73%, which validates that ECRL is able to localize the segment boundary more precisely. On Charades-STA dataset, we outperform the MGSL by 1.39%, 1.66%, 1.31% and 1.33% in all metrics, respectively. The main reasons for our proposed model outperforming the competing models lie in two folds: 1) Our newly proposed self-supervised consistency loss learns more accurate frame-wise representations for alleviating the data distribution bias, improving the generalization-ability of the model. 2) The augmented video helps the model capture the invariant query-related semantics for better predicting the transform-equivariant segment boundaries.

We further compare our method with existing works [33, 103] on the de-biased datasets [33], *i.e.*, ActivityNet-CD and Charades-CD. As shown in Table 2, our method still outperforms the other methods by a large margin.

**Comparison on efficiency.** We evaluate the efficiency of our proposed ECRL model, by fairly comparing its running time and model size in inference phase with existing methods on a single

Table 4. Main ablation study on ActivityNet dataset. Here,  $\mathcal{L}_{TSG}^1$ ,  $\mathcal{L}_{TSG}^2$  and  $\mathcal{L}_{cons}^1$ ,  $\mathcal{L}_{cons}^2$  are the grounding backbones and the consistency learning modules on the augmented and original video inputs, respectively. ‘SR’ denotes the self-refine module.

Backbone		Consistency		SR	R@1,	R@1,
$\mathcal{L}_{TSG}^1$	$\mathcal{L}_{TSG}^2$	$\mathcal{L}_{cons}^1$	$\mathcal{L}_{cons}^2$		IoU=0.5	IoU=0.7
×	✓	×	×	×	43.19	26.72
✓	✓	×	×	×	44.33	27.46
✓	✓	×	✓	×	49.28	30.01
✓	✓	✓	×	×	49.84	30.47
✓	✓	✓	✓	×	51.79	31.65
✓	✓	✓	✓	✓	<b>54.24</b>	<b>32.98</b>

Nvidia TITAN XP GPU on TACoS dataset. As shown in Table 3, it can be observed that we achieve much faster processing speeds with relatively less learnable parameters. This attributes to: 1) The proposal-based methods (ACRN, CTRL, TGN, 2DTAN, DRN) suffer from the time-consuming process of proposal generation and proposal matching. Compared to them, our grounding head is proposal-free, which is much more efficient and has less parameters. 2) The proposal-free method VLSNet utilizes convolution operation to discriminate foreground-background frames. Instead, we propose an efficient and effective consistency loss function to learn frame-wise representation.

#### 4.4 Ablation Study

We perform multiple experiments to analyze different components of our ECRL framework. Unless otherwise specified, experiments are conducted on the ActivityNet dataset.

**Main ablation.** To demonstrate the effectiveness of each component in our ECRL, we conduct ablation studies regarding the components (*i.e.*, two grounding heads on augmented video and original video in backbone model, two consistency constraints in the SSCL, and the self-refine module in the video encoder) of ECRL, and show the corresponding experimental results in Table 4. In particular, the first line represents the performance of the baseline model ( $\mathcal{L}_{TSG}^2$ ), which only train the original video-query pairs without augmented samples and consistency loss, achieving 43.19% and 26.72% in R@1, IoU=0.5 and R@1, IoU=0.7. Comparing the results in other lines of this table, we have the following observations: 1) The spatio-temporal transformation strategy constructs the augmented samples to assist the model training, which promotes the model performance (refer to line 1-2 of the table). However, the improvement is small as the consistency between augmented and original video still is not captured. 2) Both two types of consistency losses contributes a lot to the grounding performance. Specifically, each type of them is able to learn the invariant semantics between two videos for enhancing the frame-wise representation learning, thus improving performance of the transform-equivalent segment prediction (refer to line 3-4 of the table). Utilizing both of them (refer to line 5) can further boost the performance. 3) The self-refine module also contributes to the final performance (refer to line 6) by enhancing the completeness and smoothness of the discrete augmented video frames. In total, results demonstrate the effectiveness of our each component.

**Effect of the proposed consistency loss.** As shown in Table 5, we investigate the effectiveness of the proposed self-supervised consistency module. We have the following observations from these results: 1) As for consistency knowledge, directly regarding the frames as negative samples (w/o Gaussian prior) achieves worse performance than the w/ Gaussian prior variant. It indicates that the neighboring frames around the reference frame are highly correlate, a more soft gaussian-based contrastive way can lead more fine-grained representation learning in videos. Beside, the variance

Table 5. Effect of the consistency loss on ActivityNet.

Components	Changes	R@1, IoU=0.5	R@1, IoU=0.7
Consistency Knowledge	w/ Gaussian prior	<b>54.24</b>	<b>32.98</b>
	w/o Gaussian prior	50.18	30.06
	$\sigma^2=1$	52.47	31.29
	$\sigma^2=25$	<b>54.24</b>	<b>32.98</b>
Consistency Loss	$\sigma^2=100$	51.65	30.83
	w/ only $\mathcal{V}_{seg}$	52.38	31.46
	w/ $\mathcal{V}_{seg}, \mathcal{V}_{right}$	53.59	32.29
	w/ $\mathcal{V}_{left}, \mathcal{V}_{seg}$	53.64	32.31
	w/ $\mathcal{V}_{left}, \mathcal{V}_{seg}, \mathcal{V}_{right}$	<b>54.24</b>	<b>32.98</b>

Table 6. Study on temporal transformation on ActivityNet.

Components	Changes	R@1, IoU=0.5	R@1, IoU=0.7
Transform where?	only $\mathcal{V}_{seg}$	51.96	31.35
	$\mathcal{V}_{left}, \mathcal{V}_{right}$	50.78	30.42
	$\mathcal{V}_{left}, \mathcal{V}_{seg}, \mathcal{V}_{right}$	<b>54.24</b>	<b>32.98</b>
Hyper-Parameter	$\alpha = 0.7$	53.47	32.51
	$\alpha = 0.8$	<b>54.24</b>	<b>32.98</b>
	$\alpha = 0.9$	54.19	32.96

$\sigma^2$  of the prior Gaussian distribution controls how the adjacent frames are similar to the reference frame, on the assumption. It shows that too small variance ( $\sigma^2 = 1$ ) or too large variance ( $\sigma^2 = 100$ ) degrades the performance. We use  $\sigma^2 = 25$  by default. 2) As for consistency loss, we find that discriminating the frames in  $\mathcal{V}_{seg}$  already achieves very great grounding performance by learning the query-invariant semantics. Discriminating the frames in  $\mathcal{V}_{left}, \mathcal{V}_{right}$  can further boost the performance by distinguishing foreground-background.

**Study on different temporal transformations.** Here, we study the different temporal transformation, including transforming which sub-video and the sampling ratio  $\alpha$ . Table 6 shows the results. From the table, we can see that all three sub-videos are crucial for discriminating the query-relevant and query-irrelevant frame representations. Further, the whole model achieves the best performance when the sampling ratio  $\alpha$  is set to 0.8.

**Evaluation of different backbone models.** Our proposed self-supervised consistency learning can serve as a “plug-and-play” for existing TSG methods. As shown in Table 7, we demonstrate the effectiveness of our proposed method by directly applying our augmentation strategy and the SSCL to other TSG models. It shows that our method helps to learn more discriminate frame-wise features for learning query-invariant semantics and predicting transform-equivariant segment, improving the generalization-ability of the model.

#### 4.5 Visualization

As shown in Figure 4, we give the qualitative examples of the grounding results. Compared to VSLNet and MGSL, our method can learn more discriminative frame-wise representations and ground the segment more accurately.

Table 7. Evaluation of different grounding backbones. We apply our augmentation and SSCL on existing TSG models.

Methods	Changes	R@1, IoU=0.5	R@1, IoU=0.7
LGI	Original	41.51	23.07
	+ ECRL	<b>50.25</b>	<b>28.63</b>
VSLNet	Original	43.22	26.16
	+ ECRL	<b>53.48</b>	<b>33.76</b>



Fig. 4. The visualization examples of grounding results.

## 5 CONCLUSION

In this paper, we propose a novel Equivariant Consistency Regulation Learning (ECRL) framework to enhance the generalization-ability and robustness of the TSG model. Specifically, we introduce a video data augmentation strategy to construct synthetic samples, and propose a self-supervised consistency loss to learn the semantic-invariant frame-wise representations for assisting model learning and predicting transform-equivariant segment boundaries. Experimental results on three challenging benchmarks demonstrate the effectiveness of the proposed ECRL.

## REFERENCES

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*. 5803–5812.
- [2] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. 2020. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9922–9931.
- [3] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In *EMNLP*. 9810–9823.
- [4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*. 6299–6308.
- [5] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 162–171.
- [6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chile Tan, and Xiaolin Li. 2020. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *AAAI*. 10551–10558.
- [7] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. 2020. Learning Modality Interaction for Temporal Sentence Localization and Event Captioning in Videos. In *ECCV*. Springer, 333–351.
- [8] Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8199–8206.
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [10] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3584–3592.
- [11] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*. 1422–1430.

- [12] Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees GM Snoek. 2016. Early embedding and late reranking for video captioning. In *Proceedings of the 24th ACM international conference on Multimedia*. 1082–1086.
- [13] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.
- [14] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2022. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2022), 4065–4080.
- [15] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. 2022. Hierarchical Contrast for Unsupervised Skeleton-based Action Representation Learning. *arXiv preprint arXiv:2212.02082* (2022).
- [16] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2022. Multi-Modal Cross-Domain Alignment Network for Video Moment Retrieval. *IEEE Transactions on Multimedia* (2022).
- [17] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. 2023. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. *arXiv preprint arXiv:2303.07863* (2023).
- [18] Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. 2022. Hierarchical Local-Global Transformer for Temporal Sentence Grounding. *arXiv preprint arXiv:2208.14882* (2022).
- [19] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 2021. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3299–3309.
- [20] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*. 5267–5275.
- [21] Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019. Structured two-stream attention network for video question answering. In *AAAI*.
- [22] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 245–253.
- [23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv* (2018).
- [24] Jacob Goldberger, Shiri Gordon, Hayit Greenspan, et al. 2003. An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures.. In *ICCV*, Vol. 3. 487–493.
- [25] Chao Guo, Daizong Liu, and Pan Zhou. 2022. A Hybrid Alignment Loss for Temporal Moment Localization with Natural Language. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [26] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video representation learning by dense predictive coding. In *ICCV Workshops*. 0–0.
- [27] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021. Video moment localization via deep cross-modal hashing. *IEEE TIP* 30 (2021), 4667–4677.
- [28] Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wang, and Xian-Sheng Hua. 2021. Coarse-to-fine semantic alignment for cross-modal moment localization. *IEEE Transactions on Image Processing* 30 (2021), 5933–5943.
- [29] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *ECCV*. 499–515.
- [30] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*. 706–715.
- [31] Haoifei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. 2021. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3195–3204.
- [32] Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. 2022. A Closer Look at Debaised Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* (oct 2022). <https://doi.org/10.1145/3565573>
- [33] Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. 2022. A Closer Look at Debaised Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach. *arXiv* (2022).
- [34] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. 2023. A Survey on Temporal Sentence Grounding in Videos. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 19, 2, Article 51 (feb 2023), 33 pages. <https://doi.org/10.1145/3532626>
- [35] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *CVPR*. 9972–9981.
- [36] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3889–3898.
- [37] Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single shot temporal action detection. In *ACM MM*. 988–996.
- [38] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. 2023. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia* (2023).



- [39] Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. 2023. Hypotheses Tree Building for One-Shot Temporal Sentence Localization. *arXiv preprint arXiv:2301.01871* (2023).
- [40] Daizong Liu and Wei Hu. 2022. Learning to Focus on the Foreground for Temporal Sentence Grounding. In *Proceedings of the 29th International Conference on Computational Linguistics*. 5532–5541.
- [41] Daizong Liu and Wei Hu. 2022. Rethinking Graph Neural Networks for Unsupervised Video Object Segmentation. (2022).
- [42] Daizong Liu and Wei Hu. 2022. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4536–4545.
- [43] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. 2020. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing* 413 (2020), 145–157.
- [44] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1665–1673.
- [45] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2020. Reasoning Step-by-Step: Temporal Sentence Localization in Videos via Deep Rectification-Modulation Network. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1841–1851.
- [46] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. 2021. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9292–9301.
- [47] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *CVPR*. 11235–11244.
- [48] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4092–4101.
- [49] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4070–4078.
- [50] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1683–1691.
- [51] Daizong Liu, Xiaoye Qu, and Pan Zhou. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9302–9311.
- [52] Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. 2022. Exploring Motion and Appearance Information for Temporal Sentence Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [53] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. 2021. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2100–2108.
- [54] Daizong Liu, Dongdong Yu, Changhu Wang, and Pan Zhou. 2021. F2net: Learning to focus on the foreground for unsupervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2109–2117.
- [55] Daizong Liu, Hongting Zhang, and Pan Zhou. 2021. Video-based facial expression recognition using graph convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 607–614.
- [56] Daizong Liu and Pan Zhou. 2023. Jointly visual-and semantic-aware graph memory networks for temporal sentence localization in videos. *arXiv preprint arXiv:2303.01046* (2023).
- [57] Daizong Liu, Pan Zhou, Zichuan Xu, Haozhao Wang, and Ruixuan Li. 2022. Few-Shot Temporal Sentence Grounding via Memory-Guided Semantic Learning. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).
- [58] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *SIGIR*. 15–24.
- [59] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*. 289–297.
- [60] Ziyang Ma, Xianjing Han, Xuemeng Song, Yiran Cui, and Liqiang Nie. 2021. Hierarchical deep residual reasoning for temporal moment localization. In *ACM Multimedia Asia*. 1–7.
- [61] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 527–544.
- [62] Shentong Mo, Daizong Liu, and Wei Hu. 2022. Multi-Scale Self-Contrastive Learning with Hard Negative Mining for Weakly-Supervised Query-based Video Grounding. *arXiv preprint arXiv:2203.03838* (2022).

- [63] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *CVPR*. 10810–10819.
- [64] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2765–2775.
- [65] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. 2017. Representation learning by learning to count. In *Proceedings of the IEEE international conference on computer vision*. 5898–5906.
- [66] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. *arXiv* (2020).
- [67] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.
- [68] AJ Piergiovanni and Michael Ryoo. 2019. Temporal gaussian mixture layer for videos. In *International Conference on Machine Learning*. PMLR, 5152–5161.
- [69] Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. 2021. Spatial-temporal action localization with hierarchical self-attention. *IEEE Transactions on Multimedia* 24 (2021), 625–639.
- [70] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6964–6974.
- [71] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [72] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.
- [73] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*. 2464–2473.
- [74] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45 (1997), 2673–2681.
- [75] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1049–1058.
- [76] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*. 510–526.
- [77] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *CVPR*. 5179–5187.
- [78] Che Sun, Hao Song, Xinxiao Wu, Yunde Jia, and Jiebo Luo. 2021. Exploiting informative video segments for temporal action localization. *IEEE Transactions on Multimedia* 24 (2021), 274–287.
- [79] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.
- [80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- [81] Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [82] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*.
- [83] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *AAAI*.
- [84] Zeyu Xiong, Daizong Liu, and Pan Zhou. 2022. Gaussian Kernel-Based Cross Modal Network for Spatio-Temporal Video Grounding. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2481–2485.
- [85] Zeyu Xiong, Daizong Liu, Pan Zhou, and Jiahao Zhu. 2023. Tracking Objects and Activities with Attention for Temporal Sentence Grounding. *arXiv preprint arXiv:2302.10813* (2023).
- [86] Huijuan Xu, Abir Das, and Kate Saenko. 2019. Two-stream region convolutional 3D network for temporal activity detection. *IEEE transactions on pattern analysis and machine intelligence* 41, 10 (2019), 2319–2332.
- [87] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9062–9069.
- [88] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. 2020. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10156–10165.

- [89] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. 2019. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 314–323.
- [90] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. 2020. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing* 29 (2020), 8535–8548.
- [91] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1339–1348.
- [92] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.
- [93] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. 2021. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*.
- [94] Jin Yuan, Yi-Liang Zhao, Huanbo Luan, Meng Wang, and Tat-Seng Chua. 2014. Memory Recall Based Video Search: Finding Videos You Have Seen before Based on Your Memory. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 10, 2, Article 21 (2014), 21 pages. <https://doi.org/10.1145/2534409>
- [95] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. 2021. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Human-centric Multimedia Analysis*.
- [96] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NIPS*. 534–544.
- [97] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, Vol. 33. 9159–9166.
- [98] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *CVPR*. 10287–10296.
- [99] Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, and Zheng Qin. 2022. Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 18, 2, Article 56 (2022), 21 pages. <https://doi.org/10.1145/3478025>
- [100] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Nanning Zheng, and Gang Hua. 2021. Action coherence network for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia* 24 (2021), 1857–1870.
- [101] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*. 1247–1257.
- [102] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *ACL*.
- [103] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2021. Towards debiasing temporal sentence grounding in video. *arXiv* (2021).
- [104] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*. Springer, 649–666.
- [105] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.
- [106] Yaqing Zhang, Xi Li, and Zhongfei Zhang. 2019. Learning a Key-Value Memory Co-Attention Matching Network for Person Re-Identification. In *AAAI*, Vol. 33. 9235–9242.
- [107] Zongmeng Zhang, Xianjing Han, Xuemeng Song, Yan Yan, and Liqiang Nie. 2021. Multi-modal interaction graph convolutional network for temporal language localization in videos. *IEEE Transactions on Image Processing* 30 (2021), 8265–8277.
- [108] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. 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*. 655–664.
- [109] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2914–2923.
- [110] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–21.
- [111] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, et al. 2023. Rethinking the Video Sampling and Reasoning Strategies for Temporal Sentence Grounding. *arXiv preprint arXiv:2301.00514* (2023).