Self-Supervised Learning for Semi-Supervised Temporal Language Grounding

Fan Luo, Shaoxiang Chen, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang, Senior Member, IEEE

Abstract—Given a text description, Temporal Language Grounding (TLG) aims to localize temporal boundaries of the segments that contain the specified semantics in an untrimmed video. TLG is inherently a challenging task, as it requires comprehensive understanding of both sentence semantics and video contents. Previous works either tackle this task in a fully-supervised setting that requires a large amount of temporal annotations or in a weakly-supervised setting that usually cannot achieve satisfactory performance. Since manual annotations are expensive, to cope with limited annotations, we tackle TLG in a semi-supervised way by incorporating self-supervised learning, and propose Self-Supervised Semi-Supervised Temporal Language Grounding (S²TLG). S²TLG consists of two parts: (1) A pseudo label generation module that adaptively produces instant pseudo labels for unlabeled samples based on predictions from a teacher model; (2) A self-supervised feature learning module with inter-modal and intra-modal contrastive losses to learn video feature representations under the constraints of video content consistency and video-text alignment. We conduct extensive experiments on the ActivityNet-CD-OOD and Charades-CD-OOD datasets. The results demonstrate that our proposed S²TLG can achieve competitive performance compared to fully-supervised state-of-the-art methods while only requiring a small portion of temporal annotations.

Index Terms—Temporal language grounding, Semi-supervised learning, Contrastive learning.

I. INTRODUCTION

VIDEO understanding is a challenging task in computer vision. Many related tasks have been extensively studied due to its wide applications, such as video captioning [1], action localization [46] and video question answering [47]. Temporal language grounding (TLG) [2], [3], which aims at automatically locating the temporal boundaries of the segments indicated by natural language descriptions in an untrimmed video, has recently attracted increasing attention from both the computer vision and the natural language processing communities because of its potential applications in video search engines and automated video editing.

The state-of-the-art models for tackling the temporal language grounding task in a fully-supervised setting can be roughly categorized into two groups: proposal-based models [2], [4], [5], [6], [7], [14], [19] and regression-based (proposal-free) models [8], [9], [10], [11], [12], [13], [15], [16], [17], [28]. Their major difference is the way of generating temporal predictions and training/inference efficiency (e.g., memory consumption and convergence rate). But all of these methods require precise temporal boundaries of the sentence in the video for model training. Since manually annotating such temporal labels is expensive and time-consuming, it is extremely challenging to scale up such fully-supervised methods.

To alleviate the annotation cost, weakly-supervised and semi-supervised learning methods are common solutions. Weakly-supervised temporal language grounding methods require only video-sentence pairs for training, that is, without the requirement for temporal boundary annotations. However, recent studies have shown that the overall performance of weakly-supervised methods [18], [20], [21], [22], [23], [25] are greatly inferior to fully-supervised methods in temporal language grounding. Compared with weakly-supervised learning, semi-supervised learning can make use of a limited amount of annotated data as necessary supervision to achieve higher performance. Thus, semi-supervised methods could be a more appropriate solution for problems where annotations are expensive but not completely impossible to obtain. Though semi-supervised learning has been widely explored in tasks such as object detection [28], [34], temporal action localization [29], [30], to the best of our knowledge, we are the first to explore semi-supervised learning for temporal language grounding.

The major challenge in semi-supervised temporal language grounding is how to learn from the large amount of coarsely annotated data. One common practice for semi-supervised learning is generating pseudo labels, in TLG, we have to handle different forms of labels as previously mentioned, e.g. proposal scores or boundary coordinates. Besides generating pseudo labels, learning a proper feature representation is also crucial for this task and is also our focus. Recent progress of self-supervised learning suggests that preserving consistency between different views of samples (generated by perturbation/augmentation) is an effective way to make use of unlabeled data [31], [63], [64], [65], [66]. However, these methods mostly focus on the intra-modal consistency alone and can not make full use of the multi-modal (video and text) data in TLG, which could lead to sub-optimal
performance. For the TLG task, it is essential to take into consideration the rich textual information and perform inter-modal feature learning. Moreover, it is computationally expensive to perform perturbation/augmentation in the pixel space, thus we need to design proper perturbations in the feature space for this task to facilitate efficient contrastive feature learning.

The goal of this paper is to tackle the above challenges. Our proposed framework correspondingly consists of two parts: pseudo label generation and self-supervised feature learning. The pseudo-label generation aims at producing more precise pseudo labels for different types of grounding models. In the semi-supervised training process, we adopt the teacher-student scheme to use a teacher model to produce instant pseudo labels for the student model to leverage unlabeled data. The self-supervised feature learning includes intra-modal and inter-modal contrastive learning, for learning video feature representations under the constraint of video content consistency and video-text alignment, respectively. To produce different views of a sample, we design two types of sequential data perturbations: time tagging and video compression simulation, which can make the learned video features more robust and also focus more on the semantics of video content to preserve consistency within the video modality and between the video and text modalities. Our proposed framework is abbreviated as $S^4$TLG (Self-Supervised Semi-Supervised Temporal Language Grounding).

In summary, our main contributions are summarized as follows: 1) To the best of our knowledge, we are the first to study semi-supervised temporal language grounding and propose a general framework that works well for both regression-based and proposal-based grounding models. 2) We have specifically designed inter-modal and intra-modal contrastive loss functions for self-supervised feature learning in temporal language grounding, which can effectively improve grounding performance. 3) We extensively tested the proposed $S^4$TLG on two challenging out-of-distribution datasets and achieve results comparable to fully-supervised baselines for regression-based and proposal-based models.

II. RELATED WORK

Temporal language grounding aims to locate the temporal boundary of events in untrimmed videos given language queries [2], [3]. Existing methods for temporal language grounding can be mainly categorized into two groups. Proposal-based methods [2], [3], [4], [5], [6], [7], [14], [19], [48] retrieve the most confident proposal from a set of candidates. Initially, [2] follows a two-stage paradigm, which first generates candidates by temporal sliding windows and then matches language queries with these candidates. [19] designs a fine-grained and flexible 2D temporal feature map proposal generation mechanism to capture relations among temporal segments. Although those methods achieve promising performance, they are limited by the need of extra computational cost and high-quality ground-truth labels. Regression-based methods [8], [9], [10], [11], [12], [13], [15], [16], [72] are proposed to solve temporal language grounding task in an end-to-end manner. Some methods [8] focus on cross-modal attention, while others propose to learn fine-grained cross-modal interactions with a segment tree [16] or local-global context [13]. [17] attempts to address grounding in a span-based question answering way. Since annotations are expensive, significant effort has been devoted to solving temporal language grounding problem in a weakly-supervised way recently [20], [21], [22], [24], [26], [27], [71]. The mainstream approaches mostly follow the multiple instance learning paradigm [22]. Apart from this, [24] proposes to solve the weakly-supervised temporal language grounding based on contrastive learning.

Self-supervised learning (SSL) is a dominant approach in machine learning to facilitate learning with limited labeled data and large amounts of unlabeled data. SSL has been extensively studied [35], [36], [37], [38], [39], [42] in image classification. Their focus can be roughly categorized into two groups: pseudo label generation and consistency regularization. Pseudo label generation produces artificial labels with a teacher model for unlabeled data and using them to further train the student model. [39] proposes to train the student model with a fixed pre-trained teacher model. Subsequent studies investigate the usage of pseudo labels. For instance, [36] proposes Noisy-Student to train a student model with an iterative teacher model; [37] proposes FixMatch to apply different levels of augmentation to pseudo label generation and prediction; [38] updates the teacher model from the feedback of the student model. Consistency regularization methods encourage models to generate consistent outputs from perturbed input data and have been widely used in SSL. [42] maintains a regularization for ensemble prediction. [35] updates the teacher model by averaging weights of the student model at each training step so as to promote the quality of labels. Most related to our work are two previous methods that study semi-supervised temporal action localization (SSTAL). [29] designs two sequential perturbations based on the mean teacher framework, [30] proposes to incorporate self-supervised learning into SSTAL. But language grounding is much more complicated than action localization and requires inter-modal joint learning.

Self-supervised learning leverages unlabeled data to make the model learn intrinsic knowledge from data. Previous methods rely on pretext tasks to learn better feature representations—representative ones are order verification [54], sequence sorting [55], masked feature reconstruction and clip-order prediction [30]. Recently, contrastive learning methods [32], [33], [57], [58], [59], [60], [63], [64], [65], [66], have attracted increasing attention since they are proven to be able to learn more powerful feature representations for downstream tasks. Its basic idea is to align similar samples in feature space and push away the dissimilar samples, so that the learned features can be discriminative.

III. OUR APPROACH

In this section, we first give the problem definition of the semi-supervised temporal language grounding task. Then we introduce the details of our proposed $S^4$TLG framework, which includes pseudo label generation, sequential perturbation, and self-supervised feature learning.
A. Problem Formulation

Given an untrimmed video \( \mathbf{V} \in \mathbb{R}^T \) and a corresponding sentence description \( \mathbf{S} \in \mathbb{R}^N \), where \( T \) is the length of video sequence and \( N \) is the number of words, temporal language grounding aims at locating the video segment \( \tau = [s, e] \) which is mostly related to the sentence. In the semi-supervised setting, only a part of the video-sentence pairs are provided with temporal boundary ground truth while the remaining ones are not available during training. The grounding model can be formulated as:

\[
\tau = f(\mathbf{V}, \mathbf{S}),
\]

where \( f(\cdot, \cdot) \) is usually a composite of feature encoding, cross-modal processing, segment predicting, and post-processing steps. Since the focus of this paper is the learning process, so we use vanilla grounding models instead of advanced ones.

B. Pseudo Label Generation

The loss functions of temporal language grounding can take different forms, requiring different methods for generating pseudo labels. Below, we discuss how to derive pseudo labels for regression-based grounding and proposal-based grounding, respectively.

Regression-based Model. Generally, regression-based grounding models can be formulated as:

\[
\tau = f_r(\mathbf{V}, \mathbf{S}),
\]

\[
f_r(\mathbf{V}, \mathbf{S}) = \Omega_r(\Phi_{int}(\varphi_V(\mathbf{V}), \varphi_S(\mathbf{S}))),
\]

where \( \Omega_r(\cdot) \) denotes the temporal boundary regression module, \( \Phi_{int}(\cdot, \cdot) \) is the cross-modal interaction module, and \( \varphi_V(\cdot) \) and \( \varphi_S(\cdot) \) are the feature encoding modules for the video and sentence, respectively. Following previous work [8], \( \varphi_V \) is comprised of a CNN and a bi-directional LSTM unit, \( \varphi_S \) is comprised of a Glove embedding layers and a bi-directional LSTM unit. \( \Phi_{int} \) denotes an iterative cross-modal co-attention operation:

\[
h_r, a = \Phi_{int}(\varphi_V(\mathbf{V}), \varphi_S(\mathbf{S})),
\]

where \( h_r \in \mathbb{R}^{T \times D_h} \) is the attended video feature and \( a \in \mathbb{R}^T \) is the corresponding attention weights. Then the temporal segment is predicted as:

\[
\hat{\tau}_r = [s, e] = \Omega_r(h_r),
\]

where \( \Omega_r(\cdot) \) can be a multi-layer perceptron.

During training, there are two types of loss functions:

\[
\mathcal{L}_r = \mathcal{L}_r^{\text{reg}} + \alpha \mathcal{L}_r^{\text{cal}}.
\]

\( \mathcal{L}_r^{\text{reg}} \) is the temporal boundary regression loss, which is the smooth L1 loss

\[
\mathcal{L}_r^{\text{reg}}(\tau_r, \hat{\tau}_r) = \text{SmoothL1}(s, \hat{s}) + \text{SmoothL1}(e, \hat{e}),
\]

where \( \hat{\tau}_r \) is the ground truth segment label\(^1\), \( \mathcal{L}_r^{\text{cal}} \) is the attention calibration loss to encourage the alignment between the attention

\( ^1 \)We denote ground truth labels by symbols with a hat (e.g., \( \hat{\tau}_r \)), and pseudo labels by symbols with an asterisk (e.g., \( \tau^*_r \)).

weights and the ground truth \( \hat{a} \in \mathbb{R}^T \):

\[
\mathcal{L}_r^{\text{cal}}(\hat{a}, a) = -\sum_{t=1}^{T} \hat{a}_t \log a_t + \sum_{t=1}^{T} a_t,
\]

\( \hat{a}_t \) is 1 if \( s \leq t \leq e \), and 0 otherwise. We design pseudo labels for both types of losses. \( \tau^*_r \) and \( a^*_r \) correspond to ground truth labels \( \hat{\tau}_r \) and \( \hat{a} \). We directly use the teacher-model’s boundary prediction \( \hat{\tau}_r \) as the pseudo label \( \tau^*_r \), and then construct attention weight pseudo label \( a^*_r \) based on \( \tau^*_r \). At time step \( t \), the attention weight is expected to be close to 1 if \( s^* \leq t \leq e^* \), and otherwise it should be close to 0:

\[
a^*_r = \begin{cases} 1 & s^* \leq t \leq e^*, \\ 0 & \text{otherwise}. \end{cases}
\]

Proposal-based Model. The proposal-based methods can be formulated as below:

\[
\omega_p = f_p(\mathbf{V}, \mathbf{S}),
\]

\[
f_p(\mathbf{V}, \mathbf{S}) = \Omega_p(\Phi_{int}(\varphi_V(\mathbf{V}), \varphi_S(\mathbf{S}))).
\]

In contrast to regression-based methods, the output of \( f_p(\cdot, \cdot) \) are the candidate proposals, i.e., \( \omega_p \in \mathbb{R}^{T \times K} \) represents \( K \) anchors at each of the \( T \) time steps. And the final segment \( \tau_p \) is selected for \( \omega_p \) with a non-maximum suppression (NMS) process. To make this a fair comparison, the feature encoding module \( \varphi_V \) and \( \varphi_S \) are kept consistent with the above regression-based method. The cross-modal interaction module \( \Phi_{int} \) in the proposal-based model firstly fuses multi-modal feature with a linear transformation, then applies self-attention to generate the contextual integration feature \( h_p \in \mathbb{R}^{T \times D_h} \):

\[
h_p = \Phi_{int}(\mathbf{V}, \mathbf{S}).
\]

Following [48], the proposal generation module \( \Omega_p \) is a combination of an anchor prediction submodule and a boundary prediction submodule:

\[
c, b = \Omega_p(h_p),
\]

where \( c \in [0, 1]^{T \times K} \) is the confidence score of each candidate proposal, and \( b \in [0, 1]^T \) represents the probability of being a boundary point. During inference, the anchor score and boundary probability are jointly used to produce the final score of a candidate proposal. The proposal-based loss \( \mathcal{L}_p^{\text{task}} \) is then defined as below:

\[
\mathcal{L}_p^{\text{task}} = \mathcal{L}_p^{\text{anchor}} + \alpha \mathcal{L}_p^{\text{boundary}}.
\]

The anchor loss \( \mathcal{L}_p^{\text{anchor}} \) is calculated with the prediction \( c \) and label: \( \hat{c} \in \mathbb{R}^{T \times K} \)

\[
\mathcal{L}_p^{\text{anchor}}(c, \hat{c}) = -\sum_{t=1}^{T} \sum_{k=1}^{K} w_k^+ \hat{c}_{tk} \log c_{tk}
\]

\[
+ w_k^- (1 - \hat{c}_{tk}) \log (1 - c_{tk}),
\]

where \( \hat{c}_{tk} \) is 1 if the corresponding proposal has a temporal IoU with the ground truth segment \( \hat{\tau} \) larger than 0.5 and 0 otherwise, and the \( w_k^+ \), \( w_k^- \) are the positive and negative sample weights.
of the \( k \)-th proposal. The boundary loss \( \mathcal{L}^{boundary}_p \) is calculated with the label \( b \in \mathbb{R}^T \), where \( b_t = 1 \) if \( t \) equals \( s \) or \( e \), otherwise \( b_t = 0 \). \( w^+ \) and \( w^- \) are positive and negative weights. The loss is formulated as:

\[
\mathcal{L}^{boundary}_p(b, \hat{b}) = - \sum_{t=1}^{T} w^+_b \log b_t + w^-_b (1 - \hat{b}_t) \log (1 - b_t). \tag{14}
\]

Compared with the regression-based method, there are a set of candidate proposals \( \omega_i \) in the proposal-based model. We firstly select the candidate with the highest score as the pseudo label \( \tau^*_p \), then the boundary pseudo label \( b^* \) is defined as:

\[
b^*_t = \begin{cases} 1 & t = s^* \text{ or } t = e^*, \\ 0 & \text{otherwise}. \end{cases} \tag{15}
\]

As for the anchor pseudo label \( e^* \), we first compute the IoU scores \( \text{score}_{tk} \) of each anchor with the pseudo segment label \( \tau^*_p \), and then the anchor pseudo labels are assigned as

\[
e^*_{tk} = \begin{cases} 1 & \text{score}_{tk} \geq 0.5, \\ 0 & \text{score}_{tk} < 0.5. \end{cases} \tag{16}
\]

### C. Sequential Perturbation

In the literature, applying stochastic perturbations to the inputs has been proven crucial for semi-supervised models [29], [30] and is also widely adopted by (self-supervised) contrastive learning methods [32], [33]. We also consider applying sequential perturbations for both semi-supervised learning and self-supervised learning.

In contrast to simply adding gaussian noise to feature maps in Mean Teacher [35], we want to further utilize the inherent property of sequence data (i.e., video). Intuitively, due to the redundancy and temporal semantic consistency of video, moderate sequential perturbations such as short-term lagging and lossy compression would not change the semantics of a video. For example, people can hardly feel the difference between the original video and the video with lossy compression or occasional lagging, and can still obtain the same semantic information. Based on this insight, for the sequential video feature \( V \in \mathbb{R}^{T \times D} \), we design two essential types of sequential perturbations: \(^2\) time lagging and video compression simulation.

**Time lagging.** As shown in Fig. 1, the time lagging perturbation performs random temporal disturbance to a feature sequence. Given a sequential video feature \( V \), we randomly select a small segment \( r = [r_s, r_e] \), \( 1 \leq r_s \leq r_e \leq T \), \( \frac{r_e - r_s}{(r_e - r_s)} \leq 4\% \) and randomly shuffle the features along the channel dimension within the segment. We keep the channel permutation temporally consistent to avoid overly perturbing the semantics. Previous work [30] performs temporal feature flipping which flips the whole feature sequence along the temporal dimension. The difference between temporal feature flipping and time lagging is that temporal feature flipping reverses the whole feature while the time lagging randomly selects a small local segment and shuffles those features along the channel dimension. Thus time lagging retains more semantic information and temporal structure than temporal feature flipping. [30] is applied to temporal localization of actions, which is usually less complex and less sensitive to temporal flipping. In the temporal language ground-task, however, the temporal structure is critical for localizing the sentence and flipping the whole video sequence can dramatically change the semantics of a video. Thus our designed time lagging is a more suitable perturbation for TLG.

**Video compression simulation.** We also simulate lossy video compression in the feature space as a type of perturbation since it

\(^2\)The perturbations are applied to the video features instead of the video frames, since the latter makes the computational cost much higher.
generally preserves the essential contents. Inspired by the group of pictures structure (I, B, P frames) in video compression standards [61], we define three types of features that are constructed differently. As shown in 1, given a target compression rate \( \theta \), I-feature is obtained from the original video feature by uniform sampling with step size \( \frac{1}{\theta} \), and is used as the reference feature afterwards. P-feature is obtained by mean pooling the two preceding reference features, and B-feature is obtained by the mean pooling of the nearest two reference features before and after it. The reference features for P-feature is I-feature, and the referenced features for B-feature can be I-feature or P-feature. The resulting features can be regarded as video event representation with different compression rates but the same semantics, and increasing the compression rate leads to more attention being paid to the overall video semantics.

As shown in Fig. 1, during semi-supervised learning, the sequential perturbation is used to augment the input data for both teacher and student models. Meanwhile, in self-supervised learning (Sec III-D), sequential perturbations are also used to generate different views for given input videos, forcing the model to learn the underlying semantic features and be invariant to these perturbations.

D. Self-Supervised Feature Learning

We perform self-supervised feature learning on the encoded visual and textual features. Based on the previous two types of sequential perturbations (data augmentation), we design inter-modal and intra-modal contrastive losses to assist the \( S^4 \)TLG to learn discriminative multi-modal feature representation.

We first give a general formulation of contrastive loss:

\[
L_{CL}(X, Y, X^-, Y^-) = \sum_{X \in X} \max [0, \Delta - l(X, Y) + l(X^-, Y^-)] \\
+ \sum_{Y \in Y^\circ} \max [0, \Delta - l(X, Y) + l(X, Y^-)],
\]

where \( l(\cdot, \cdot) \) is a similarity function (e.g. cosine similarity), \( X, Y \in \mathbb{R}^{D_h} \) is a pair of positive samples, and \( X^- \) and \( Y^- \) are the sets of chosen negative sample for \( X \) and \( Y \), respectively. \( \Delta \) is the margin. The goal of \( L_{CL} \) is to pull positive pairs closer and push negative pairs further in the embedding space.

**Inter-modal contrastive loss.** Consider a mini batch \( \{V_i, S_i\}_{i=1}^B \) with B video and sentence features, the inter-modal contrastive loss is:

\[
L_{CL}^{inter} = \sum_{i=1}^B L_{CL}(V_i, S_i, V_i^-, S_i^-),
\]

where \( V_i^- = \{V_j | j \in [1, B] \text{ and } j \neq i\} \) and \( S_i^- = \{S_j | j \in [1, B] \text{ and } j \neq i\} \) are the negative samples for the \( i \)-th video and sentence pair. The inter-modal contrastive loss encourages the semantically related video and sentence to be similar, and decreases the similarity between semantically irrelevant video sentence pairs.

**Intra-modal contrastive loss.** To encourage the model to focus more on the semantics of the video, we also apply contrastive learning on augmented versions of video features.

\[
A_{i1}, A_{i2} = \text{Aug}(V_i),
\]

\[
L_{CL}^{intra} = \sum_{i=1}^B L_{CL}(A_{i1}, A_{i2}, A_{i1}^-, A_{i2}^-),
\]

where \( \text{Aug}(\cdot) \) takes a video sample and augments it twice using perturbations chosen randomly from the set \{time lagging, video compression simulation, identity mapping\}, and \( A_{i1}, A_{i2} \in \mathbb{R}^{T \times D_h} \) are the augmented video features. Similar to inter-modal contrastive loss, intra-modal contrastive loss also regards the rest of \( B - 1 \) samples as negative, that is \( A_{i1} = \{A_{j1} | j \in [1, B] \text{ and } j \neq i\} \) and \( A_{i2} = \{A_{j2} | j \in [1, B] \text{ and } j \neq i\} \) are the negative samples for sample \( i \).

The overall self-supervised loss is:

\[
L_{self} = L_{CL}^{inter} + L_{CL}^{intra}.
\]

E. The Training Process

There are two phases in our grounding model during the training process. In the first phase, we pre-train the model with a limited amount of annotated data in a supervised manner to obtain a teacher model. In the second phase, the teacher model generates pseudo labels for the unlabeled data in each epoch. Following the instant-teaching [28] framework, we use instant pseudo labeling in our \( S^4 \)TLG. We then perform the semi-supervised training by optimizing the semi-supervised loss \( L_{task} \) and self-supervised loss \( L_{self} \) using all the training samples:

\[
L_{all} = L_{task} + \beta L_{self},
\]

where \( L_{task} \) can be \( L_{r_{task}} \) or \( L_{p_{task}} \), depending on the type of grounding model being trained, and \( \beta \) is a hyper parameter that controls the weight of self-supervised loss.

IV. EXPERIMENTS

A. Datasets

Recent studies [44], [62] have suggested that the data distribution of popular datasets used by current TLG methods are mostly biased and can not reflect the actual grounding capabilities of state-of-the-art models. This problem can have even greater negative effects in our semi-supervised setting. Thus we decide to follow the settings in [44] and use a novel split of existing datasets, in which the training and testing data are designed to have different distributions, so that the model must learn intrinsic clues from the video and sentence.

**Charades-CD-OOD.** Charades-CD-OOD is built on the Charades-STA dataset [2], which annotates sentence descriptions in a semi-automatic way. It contains 9,848 videos and 16,128 sentence-moment pairs. On average, the video and sentence length on the Charades-CD dataset are 29.8 seconds and 8.6 words respectively. [44] propose to split sentence-moment pairs into training, validation, and test-ood set of sizes 11,071, 859 and 3,375, respectively.

**ActivityNet-CD-OOD.** ActivityNet-CD-OOD is built on the ActivityNet Captions [43] dataset, which is originally for dense video captioning. The dataset has 19,970 videos and 17,031

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
sentence-moment pairs. Compared with Charades-STA, its variation of temporal segment lengths is much larger, ranging from a few seconds to several minutes. The average sentence length is 13.2 words, and the average video length is 117.74 seconds. According to [44], ActivityNet-CD-OOD contains 51,415, 3,321, 13,578 sentence-moment pairs for training, validation, and test-ood respectively.

The rate $\psi$, which represents the percentage of labeled data in training data, is set to 5% in ActivityNet-CD-OOD and 30% in Charades-CD-OOD, to ensure that these two datasets have similar amounts of labeled instances.

### B. Metrics

Following the previous setting [44], we use the improved “dR@n, IoU@m” as the evaluation metric:

$$dR@n, IoU@m = R@n, IoU@m \cdot \gamma_s \cdot \gamma_e,$$

where the “R@n, IoU@m” is the percentage of the query sentences that our model can produce at least one appropriate moment whose IoU with the ground truth is larger than $m$ in the top-n results, and $\gamma_s = 1 - ||s - \hat{s}||$, $\gamma_e = 1 - ||e - \hat{e}||$ are coefficients used to discount the boundary offsets. Both $\tau$ and $\hat{r}$ are normalized to [0,1] by dividing the whole video length. Specifically, we report the results for $n \in \{1, 5\}$ and $m \in \{0.3, 0.5, 0.7\}$ on the Charades-CD-OOD and ActivityNet-CD-OOD datasets.

### C. Implementation Details

We extract the C3D [51] visual features on both datasets. The description sentences are encoded by 300 d GloVe [53] word embedding vectors. The video length $T$ is set as 128 for computing convenience, thus the video feature sequence will be uniformly sampled if length is greater than 128, otherwise zero-padded. The batch size is set to 32, in which the ratio of labeled data is $\psi$. Adam [52] is used as the optimizer, with a learning rate 0.001 for both regression-based method and proposal-based method. The dimension of all hidden layers is 512. Weights $\alpha_r$ and $\alpha_p$ are 0.01 and 1.0 respectively, and weight $\beta$ is 1.0. The simulated compression rate $\theta$ is randomly chosen from $\{0.25, 0.5, 0.75\}$ and margin $\Delta$ is 1.0. We take the mean of multiple experiments with different random seeds as the results.

### D. Performance Comparison

We first compare our S$^4$TLG with state-of-the-art temporal language grounding methods on both ActivityNet-CD-OOD and Charades-CD-OOD datasets. These SOTA methods can be categorized into several groups: (1) Regression-based methods: ACRN [6]: a memory-based attentive cross-modal retrieval network; ABLR [8]: an iterative attention-based location regression network; TSP-PRL [50]: a coarse-to-fine tree-structured policy based progressive reinforcement learning model; DRN [9]: a dense regression network with multiple regression loss. (2) Proposal-based methods: CTRL [2]: Cross-modal semantics alignment temporal localizer, SCDM [49]: Language guided semantic conditioned dynamic modulation network, 2D-TAN [19]: 2D temporal adjacent semantic relation encoded network, MS-2D-TAN [68]: Multi-scale 2D temporal adjacent semantic relation encoded network, RaNet [69]: Relation-aware Network, TNC-DCM [70]: Causality-inspired deconfounded video moment retrieval network. (3) Weakly-supervised method: WSSL [20]: a weakly-supervised crossing attention multi-model localization network. Note that the performance of these methods are reported by [44], these methods will mostly experience a significant performance drop on the out-of-distribution test set compared to the original test set.

**Comparisons with semi-supervised baseline.** To demonstrate the effectiveness of our proposed pseudo label generation and self-supervised feature learning, we first compare our full method with a baseline method, which is obtained by removing the label generation and self-supervised feature learning from our method and is only trained with limited labeled data (baseline$@\psi$). To demonstrate that the proposed pseudo label generation and self-supervised feature learning can be broadly applied to different types of models, we integrate them into both regression-based and proposal-based models. For both regression-based and proposal-based models, the baselines have the same model architecture with our full model $S^4$TLG. When the percentage of labeled data is $\psi$ (e.g., 5%), the baseline$@\psi$ represents the model trained with these labeled data only, and no additional unlabeled data is used. Baseline$@100\%$ for a particular $\psi$ is obtained by further training the baseline$@\psi$ with all the labeled data. Table I summarizes the performance. It can be observed that our $S^4$TLG has a significant improvement compared with the baseline for both regression-based and proposal-based models, highlighting that our proposed components can effectively utilize the coarsely annotated data to improve grounding performance. Especially for the regression-based model, our $S^4$TLG boosts the baseline from 5.62 to 8.16 and from 6.26 to 9.15 in terms of dR@1, IoU@0.7 on ActivityNet-CD-OOD and Charades-CD-OOD, respectively. Meanwhile, $S^4$TLG achieves comparable results to the baseline that uses all labeled data (baseline$@100\%$) for both types of models. The results indicate that $S^4$TLG can effectively utilize unlabeled data for learning better features via contrastive losses and generating high-quality pseudo labels. In addition, on both datasets, the regression-based model achieves more significant improvements than proposal-based model. One possible reason is that the proposal-based model relies more on the quality of labels and predefined anchor scales, thus is more sensitive to the noise in pseudo labels. The results ($S^4$TLG@100%) again indicate the effectiveness of our method even though pseudo label generation is removed (all samples have labels).

**Comparisons with SOTA methods.** We then compare our $S^4$TLG with fully- and weakly-supervised state-of-the-art methods. As Table I shows, while using limited labeled data, our $S^4$TLG significantly outperforms the weakly-supervised method WSSL [20] in all metrics. Compared with fully-supervised methods, our regression-based model can achieve competitive performance on ActivityNet-CD-OOD and Charades-CD-OOD with only 5% and 30% data labeled with temporal boundary. Our performance of $S^4$TLG@100% is quite weaker than DRN, we think there are several reasons: (1) The
SOTA models usually use complex video and sentence encoders and multiple levels of cross-modal interaction to achieve higher performances. S$^4$TLG has much simpler structure and less parameters comparing with DRN, so the DRN could have better feature learned in theory; (2) There are varying degrees of prior assumptions about the datasets within the SOTA methods. The less prior assumptions, the method is more stable with data distribution changing. Comparing with other complex grounding model, there are several methods, e.g. DRN and SCDM, have more refined designs to avoid invoking too much prior assumptions about datasets thus they could achieve much better performance. While some proposal-based methods like 2D-TAN, which rely on the prior assumption to define parameters about proposal scales, are more sensitive to distribution changing.

**Comparisons with SOTA models using limited labels.** We also compare our S$^4$TLG with SOTA models [68], [69], [70] under limited labels. As shown in Table I, our S$^4$TLG with rather simple encoders can outperform SOTA models on Charades-CD-OOD and ActivityNet-CD-OOD.

**E. Ablation Study**

To show the effectiveness of each component, we conduct ablation studies for both regression-based and proposal-based models on ActivityNet-CD-OOD dataset. The results are shown in Tables II and III.

**Pseudo Label Generation.** The performance of S$^4$TLG for regression-based model with and without pseudo label generation is shown in Table II (the 1st and 2nd row). It can be observed that the pseudo label generation improves the temporal grounding accuracy by using instant pseudo labels. The improvements demonstrate the effectiveness of our pseudo label generating module. In Table III (the 1st and 2nd row) we also observe that using pseudo labels leads to performance improvement on dR@1,IoU@0.1, but the improvements are smaller than for regression-based models. We conjecture that the reason could be that proposal-based pseudo labels suffer from noise, so it is natural to introduce extra regularization to improve the teacher model and the label quality as well.
Sequential Perturbations. Our designed perturbations, time lagging and video compression simulation, are motivated by the insight that moderate perturbation is beneficial for representation learning. To investigate their effectiveness, we compare the S^4TLG’s performance with and without sequential perturbations. The results are shown in Tables II and III (the 2\textsuperscript{nd} and 3\textsuperscript{rd} row). It is clear that the performance is improved by applying sequential perturbations to augment the training samples, and this validates that slightly altering the temporal structure of videos can help our model focus more on the semantics. We further investigate the effectiveness of individual sequential perturbation. As shown in Table IV, both perturbations can improve the performance, and time lagging is slightly more effective than video compression simulation.

Inter-modal Contrastive Loss. In order to validate the effectiveness of the inter-modal contrastive loss, we evaluate the S^4TLG with and without inter-modal contrastive loss. By comparing the 2\textsuperscript{nd} and 5\textsuperscript{th} row in both Tables II and III, we observe a significant performance improvement for all metrics, e.g., the dR@1, IoU@0.7 are improved by 10.7% and 14% for regression-based and proposal-based models, respectively. Same conclusion can be obtained by comparing the 3\textsuperscript{rd} and 6\textsuperscript{th} in Tables II and III. The results suggest that the inter-modal contrastive loss can help to learn more discriminative visual and language features. As we suggested earlier, inter-modal self-supervised learning can be used as an implicit regularization to improve the label quality especially for the proposal-based model, and the results here also agree with our previous conjecture.

F. Qualitative Results

Intra-modal Contrastive Loss. We further investigate the effect of intra-modal contrastive loss. As shown in Tables II and III (the 3\textsuperscript{rd} and 4\textsuperscript{th} row), when using intra-modal contrastive loss alone, the performance improves slightly. However, a substantial performance improvement can be obtained by combining intra-modal contrastive loss with inter-modal contrastive loss (the 6\textsuperscript{th} and 7\textsuperscript{th} row). The results suggest intra-modal contrastive loss plays a supplementary role to inter-modal contrastive loss in feature learning for temporal visual grounding task. Moreover, learning a better feature space that aligns visual and textual features well is the primary requirement for cross-modal tasks, and it can be the basis for further improving the feature learning within a particular modality.

Different Percentages of Labeled Data. We also evaluate the S^4TLG model with different percentages of labeled data. Table V shows the performance of our S^4TLG model on ActivityNet-CD-OOD and Charades-CD-OOD with different percentages of labeled training data. From the results, we can observe that S^4TLG consistently performs better than the baseline under different percentages of labeled training models, with regarding to different models on both datasets. It is worthwhile mentioning that the performance gaps between our S^4TLG and baseline@100% are small even only 5% labeled data is available for our model. Besides, for some cases, our S^4TLG performs even slightly better baseline@100%. The results validate the effectiveness of our proposed semi-supervised training and the self-supervised feature learning strategies. The performance of baseline@100% varies across different percentages of label training data. This is because baseline@100% is first pre-trained with ψ labeled data, and for fair comparison, this pre-training is kept the same as our S^4TLG. Meanwhile, it can be observed that as the amount of labeled data increases, the performance of S^4TLG quickly saturates, e.g., proposal-based S^4TLG achieves similar results for 30% and 50% labeled data on ActivityNet-CD-OOD. This also indicates that the temporal language grounding task is actually very challenging, especially on the ood datasets we used, hence we need to focus on learning intrinsic features from the video and sentence rather than overfitting to the label distribution with complex models.

Fig. 2 shows the qualitative results of regression-based S^4TLG on ActivityNet-CD-OOD. Our S^4TLG can achieve precise grounding prediction. But it still has some limitations, for the second query, the predict is a few wider than the ground truth. The cause might be our proposal method is training any form of hallucination.
Ablation Study of Different Percentages of Labeled Training Data on ActivityNet-CD-OOD and Charades-CD-OOD. C3D Features are Used in the Comparison. The baseline@100% Represents the Model Pre-Training on 100% Labeled Data

| Methods           | ActivityNet-CD-OOD | Charades-CD-OOD |
|-------------------|--------------------|-----------------|
|                   | dR@1, dR@1, dR@1, | dR@1, dR@1, dR@1, |
|                   | IoU@0.3, IoU@0.5, | IoU@0.3, IoU@0.5, |
|                   | IoU@0.7, IoU@0.7, | IoU@0.7, IoU@0.7, |
| baseline@5%       | 27.56 15.41 5.62 | 7.18 3.31 1.42 |
| S4TLG@5%          | 32.40 18.70 8.16 | 30.94 18.13 8.09 |
| baseline@100%     | 35.94 20.52 8.57 | 31.65 18.32 8.23 |
| regression-based  |                    |                 |
| baseline@30%      | 35.38 19.93 8.48 | 40.99 21.02 6.26 |
| S4TLG@30%         | 35.87 20.38 8.69 | 43.48 28.26 9.15 |
| baseline@100%     | 37.08 21.39 9.31 | 44.63 28.59 9.48 |
| baseline@50%      | 36.83 21.32 9.37 | 42.60 27.67 8.44 |
| S4TLG@50%         | 37.07 21.36 9.59 | 41.63 25.35 11.29 |
| baseline@100%     | 37.50 21.86 9.78 | 47.57 29.20 11.30 |
| proposal-based    |                    |                 |
| baseline@30%      | 28.91 15.44 6.16 | 36.00 20.91 10.33 |
| S4TLG@30%         | 31.71 18.00 8.08 | 36.14 21.16 11.21 |
| baseline@100%     | 34.09 19.02 8.03 | 39.99 23.54 11.17 |
| baseline@50%      | 35.27 19.64 8.06 | 38.62 21.99 10.54 |
| S4TLG@50%         | 35.06 19.50 8.31 | 40.72 24.07 12.11 |
| baseline@100%     | 35.32 20.10 8.54 | 40.54 23.89 11.21 |
| baseline@100%     | 35.72 20.45 8.73 | 43.46 25.88 12.88 |

**Table V**

**V. CONCLUSION**

In this paper, we introduced a semi-supervised temporal language grounding framework assisted by contrastive self-supervised learning. We designed pseudo label generation mechanisms that fit different types of grounding models and can be used in an instant updating manner during semi-supervised learning. We also presented two effective sequential perturbations, and associated with contrastive self-supervised learning, which can further boost the performance of S4TLG. Our experiments show that S4TLG achieves competitive performance compared with fully-supervised methods using a limited amount of temporal annotations, and S4TLG is universally effective for different types of temporal language grounding models.

**REFERENCES**

[1] S. Chen, W. Jiang, W. Liu, and Y. Jiang, “Learning modality interaction for temporal sentence localization and event captioning in videos,” in Proc. IEEE Eur. Conf. Comput. Vis., vol. 12349, 2020, pp. 333–351.

[2] J. Gao, C. Sun, Z. Yang, and R. Nevatia, “TALL: Temporal activity localization via language query,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 5277–5285.
[3] L. A. Hendricks et al., “Localizing moments in video with natural language,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5804–5813.

[4] R. Ge, J. Gao, K. Chen, and R. Nevatia, “MAC: Mining activity concepts for language-based temporal localization,” in *Proc. Winter Conf. Appl. Comput. Vis.*, 2019.

[5] B. Jiang, X. Huang, C. Yang, and J. Yuan, “Cross-modal video moment retrieval with spatial and language-temporal attention,” in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 217–225.

[6] M. Liu et al., “Attentive moment retrieval in videos,” in *Proc. 41st ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 15–24.

[7] Z. Liu and T. Dinh, “Temporal moment localization in videos,” in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 843–851.

[8] Y. Yuan, T. Mei, and W. Zhu, “To find where you talk: Temporal sentence localization in video with attention based location regression,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9159–9166.

[9] R. Zeng et al., “Dense regression network for video grounding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10284–10293.

[10] C. R. Opazo, E. Marrere-Taylor, F. S. Saleh, H. Li, and S. Gould, “Proposal-free temporal moment localization of a natural-language query in video using guided attention,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2453–2462.

[11] L. Chen et al., “Rethinking the bottom-up framework for query-based video localization,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10551–10558.

[12] W. Wang, Y. Huang, and L. Wang, “Language-driven temporal activity localization: A semantic matching reinforcement learning model,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 334–343.

[13] J. Mun, M. Cho, and B. Han, “Local-global video-text interactions for temporal grounding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4197–4206.

[14] D. Liu et al., “Context-aware biaffine local network for temporal sentence grounding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11235–11244.

[15] G. Nan et al., “Interventional video grounding with dual contrastive learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2765–2775.

[16] Y. Zhao, Z. Zhao, Z. Zhang, and Z. Lin, “Cascaded prediction network via segment tree for temporal video grounding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4197–4206.

[17] H. Zhang et al., “Embracing uncertainty: Decoupling and de-bias for robust temporal grounding,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–8.

[18] X. Wang et al., “Self-supervised learning for semi-supervised temporal activity understanding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10870–10876.

[19] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, “Meta pseudo labels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11557–11568.

[20] E. Araujo, D. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-labeling and confirmation bias in deep semi-supervised learning,” in *Proc. IEEE Int. Conf. Neural Net. Work.*, 2020, pp. 1–8.

[21] J. Jeong, S. Lee, J. Kim, and N. Kwak, “Consistency-based semi-supervised learning for object detection,” in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 10758–10767. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/9f0d4dae80c3d0277922f371f587292-Abstract.html

[22] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised learning results,” in *Proc. Neural Inf. Process. Syst.*, 2017. [Online]. Available: https://openreview.net/forum?id=rky8riv1rl

[23] Q. Xie, M. Luong, E. H. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10684–10695.

[24] K. Sohn et al., “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” in *Proc. Neural Inf. Process. Syst.*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/0f4d4dae80c3d0277922f371f587292-Abstract.html

[25] K. Sohn et al., “A simple semi-supervised learning framework for object detection,” in *Proc. Conf. Artif. Intell.*, 2020. [Online]. Available: https://arxiv.org/abs/2005.04757

[26] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1foOfiQe

[27] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.

[28] Y. Yuan, X. Lan, X. Wang, L. Chen, Z. Wang, and W. Zhu, “A closer look at temporal sentence grounding in videos: Datasets and metrics,” in *Proc. 2nd Int. Workshop Hum.-Centric Multimedia Anal.*, 2021, pp. 13–21.

[29] G. A. Sigurdsson et al., “Hollywood in homes: Crowdsourcing data collection for activity understanding,” in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2016, vol. 9905, pp. 510–526. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_31

[30] Z. Shou, D. Wang, and S. Chang, “Temporal action localization in untrimmed videos via multi-stage CNNs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1049–1058.

[31] M. Tapaswi et al., “MovieQA: Understanding stories in movies through question-answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4631–4640.

[32] J. Wang, L. Ma, and W. Jiang, “Temporally grounding language queries in videos by contextual boundary-aware prediction,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12168–12175. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/10897

[33] Y. Yuan, L. Ma, J. Jiang, S. Liu, and W. Zhu, “Semantic conditioned dynamic modulation for temporal sentence grounding in videos,” in *Proc. Adv. Neural Inf. Process. Syst.*, H. M. Wallach et al., Eds., 2019, pp. 217–225.

[34] Z. Shou, D. Wang, and S. Chang, “Temporal action localization in untrimmed videos via multi-stage CNNs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1049–1058.

[35] M. Tapaswi et al., “MovieQA: Understanding stories in movies through question-answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4631–4640.

[36] J. Wang, L. Ma, and W. Jiang, “Temporally grounding language queries in videos by contextual boundary-aware prediction,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12168–12175. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/10897

[37] Y. Yuan, L. Ma, J. Jiang, S. Liu, and W. Zhu, “Semantic conditioned dynamic modulation for temporal sentence grounding in videos,” in *Proc. Adv. Neural Inf. Process. Syst.*, H. M. Wallach et al., Eds., 2019, pp. 534–544. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/6863966df8194afaa29be2d2c386fb-Absract.html

[38] J. Wu, G. Li, S. Liu, and L. Lin, “Tree-structured policy based progressive refinement learning for temporally language grounding in video,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12386–12393. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6924
[51] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2015, pp. 4489–4497.

[52] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[53] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., A. Moschitti, B. Pang, and W. Daelemans, Eds. 2014, pp. 1532–1543.

[54] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: Unsupervised learning using temporal order verification,” in Proc. IEEE Eur. Conf. Comput. Vis., B. J. Leibe, N. Matas Sebe, and M. Welling, Eds., vol. 9905, 2016, pp. 527–544. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_32

[55] H. Lee, J. Huang, M. Singh, and M. Yang, “Unsupervised representation learning by sorting sequences,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 667–676.

[56] G. Gong, X. Wang, Y. Mu, and Q. Tian, “Learning temporal co-attention models for unsupervised video action localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9816–9825.

[57] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1597–1607.

[58] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, “Big self-supervised models are strong semi-supervised learners,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 22243–22255.

[59] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9729–9738.

[60] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” 2020, arXiv:2005.04297

[61] J. Watkinson, The MPEG Handbook. London, U.K., Routledge, 2012.

[62] M. Otani, Y. Nakahima, E. Rahtu, and J. Heikkila, “Uncovering hidden challenges in query-based video moment retrieval,” in Proc. Brit. Mach. Vis. Conf., 2020.

[63] R. Li et al., “Motion-focused contrastive learning of video representations,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 2105–2114.

[64] N. Behrmann, M. Fayyaz, J. Gall, and M. Norouzi, “Long short view feature decomposition via contrastive video representation learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9244–9253.

[65] J. Wang et al., “Removing the background by adding the background: Towards background robust self-supervised video representation learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11804–11813.

[66] M. Patrick et al., “Space-time crop & attend: Improving cross-modal video representation learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10540–10552.

[67] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, “Towards debiasing temporal sentence grounding in video,” 2021, arXiv:2111.04321

[68] S. Zhang, H. Peng, J. Fu, Y. Lu, and J. Luo, “Multi-scale 2D temporal adjacent networks for moment localization with natural language,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 9073–9087, Dec. 2021.

[69] J. Gao, X. Sun, M. Xu, X. Zhou, and B. Ghanem, “Relation-aware video reading comprehension for temporal language grounding,” in Proc. 2021 Conf. Empirical Methods Natural Lang. Process., 2021, pp. 3978–3988.

[70] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, “Deconfounded video moment retrieval with causal intervention,” in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2021, pp. 1–10. [Online]. Available: https://doi.org/10.1145/3404835.3462823

[71] Y. Wang, J. Deng, W. Zhou, and H. Li, “Weakly supervised temporal adjacent network for language grounding,” IEEE Trans. Multimedia, vol. 24, pp. 3276–3286, 2022.

[72] H. Tang, J. Zhu, M. Liu, Z. Gao, and Z. Cheng, “Frame-wise cross-modal matching for video moment retrieval,” IEEE Trans. Multimedia, vol. 24, pp. 1338–1349, 2022.

Fan Luo received the B.E. degree from the South China University of Technology, Guangzhou, China, in 2020. He is currently working toward the M.Eng. degree in computer science with the Academy for Engineering and Technology, Fudan University, Shanghai, China. His research interests include temporal language grounding and video retrieval.

Shaoxiang Chen received the B.S. and Ph.D. degree from the School of Computer Science, Fudan University, Shanghai, China. His research interests include multimedia and deep learning, with respect to video captioning and temporal sentence localization in videos.

Jingjing Chen received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2018. He is currently an Associate Professor with the School of Computer Science, Fudan University, Shanghai, China. Before joining Fudan University, she was a Postdoc Research Fellow with the School of Computing, National University of Singapore, Singapore. Her research interests include robust AI, multimedia content analysis, and deep learning.

Zuxuan Wu received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, with Prof. Larry Davis in 2020. He is currently an Associate Professor with the School of Computer Science, Fudan University, Shanghai, China. His research interests include computer vision and deep learning. His work has been recognized by an AI 2000 Most Influential Scholars Honorable Mention in 2021, a Microsoft Research Ph.D. Fellowship (10 people Worldwide) in 2019, and Snap Ph.D. Fellowship (10 people Worldwide) in 2017.

Yu-Gang Jiang (Senior Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2009. He is currently a Professor of Computer Science with Fudan University, Shanghai, China. During 2009–2011, he was a Postdoctoral Research Scientist with Columbia University, New York, NY, USA. His research interests include multimedia, computer vision, and robust and trustworthy AI. He was the recipient of many awards, including the inaugural ACM China Rising Star Award, 2015 ACM SIGMM Rising Star Award, Research Award for Excellent Young Scholars from NSF China, and Chang Jiang Distinguished Professorship appointed by Ministry of Education of China.