Siamese Learning with Joint Alignment and Regression for Weakly-Supervised Video Paragraph Grounding

Chaolei Tan[1] Jianhuang Lai[1,2,3] Wei-Shi Zheng[1,2,3] Jian-Fang Hu[1,2,3*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
tanchlei@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn, wszheng@ieee.org, hujf5@mail.sysu.edu.cn

Abstract

Video Paragraph Grounding (VPG) is an emerging task in video-language understanding, which aims at localizing multiple sentences with semantic relations and temporal order from an untrimmed video. However, existing VPG approaches are heavily reliant on a considerable number of temporal labels that are laborious and time-consuming to acquire. In this work, we introduce and explore Weakly-Supervised Video Paragraph Grounding (WSVPG) to eliminate the need of temporal annotations. Different from previous weakly-supervised grounding frameworks based on multiple instance learning or reconstruction learning for two-stage candidate ranking, we propose a novel siamese learning framework that jointly learns the cross-modal feature alignment and temporal coordinate regression without timestamp labels to achieve concise one-stage localization for WSVPG. Specifically, we devise a Siamese Grounding TRansformer (SiamGTR) consisting of two weight-sharing branches for learning complementary supervision. An Augmentation Branch is utilized for directly regressing the temporal boundaries of a complete paragraph within a pseudo video, and an Inference Branch is designed to capture the order-guided feature correspondence for localizing multiple sentences in a normal video. We demonstrate by extensive experiments that our paradigm has superior practicability and flexibility to achieve efficient weakly-supervised or semi-supervised learning, outperforming state-of-the-art methods trained with the same or stronger supervision.

1. Introduction

Natural Language Video Grounding (NLVG) is an essential area in vision-language understanding, which has received increasing attention due to its wide range of real-world applications such as video retrieval [3, 17, 21, 23, 80, 94], video summarization [50, 55, 59, 60, 83, 99], action segmentation [19, 27, 35, 37, 45, 72], video question answering [30, 36, 38, 41, 82, 89], etc. Most previous works focus on tackling Video Sentence Grounding (VSG) proposed in [1, 22], which targets at localizing the temporal boundaries of an individual sentence from an untrimmed video. However, localizing single sentences can be ambiguous because the contextual information conveyed by multiple sentences is necessary to uniquely determine the temporal locations of input queries. To alleviate such issue, Bao *et al*. [4] proposed to contextualize video grounding into localizing multiple events indicated by sentences of a paragraph from the video, which is called Video Paragraph Grounding (VPG).

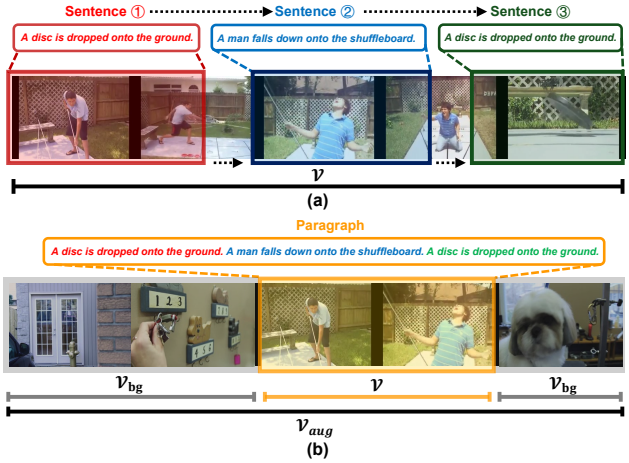


Figure 1. (a) Chronological cross-modal alignment in a video and its paired sentences. (b) Pseudo boundary supervision for regressing paragraph timestamps in a composed pseudo video.

Although remarkable progress has been attained in tackling VSG and VPG problems under a fully-supervised setting, the extremely prohibitive overheads of manually annotating temporal boundaries for language queries limit these methods to utilize large-scale video-text data. In addition, the subjectivity of annotators inevitably brings label noises that may adversely affect model training. Recently, weakly-supervised methods free of temporal annotations have become increasingly popular in the area of video grounding, aiming to address the above limitations. These approaches can be mainly categorized into multiple instance learning methods and reconstruction learning methods and are typ-

*Corresponding author

ically based on a propose-and-rank pipeline. Nevertheless, both of these paradigms assume the contributions of proposals to the contrastive or reconstruction loss accurately represent the proposal quality, which is not necessarily the case during model learning. Moreover, the quadratic complexity of proposal schemes prevents them from scaling up in parameters or training data, and the adopted supervision of video-level contrast or query-based reconstruction in prior works is not temporal-sensitive thus suffering from a huge gap with fully-supervised temporal guidance. Besides, all of the weakly-supervised methods in video grounding are tailored for tackling VSG while the weakly-supervised setting of VPG (i.e., WSVPG) has been understudied so far.

To circumvent the aforementioned drawbacks and explore an efficient weakly-supervised framework for VPG, we seek to mine the unique characteristics and underlying supervision from the intrinsic structure of video-paragraph pairs for model training. On the one hand, as observed in Figure 1 (a), the temporal location of an event is highly correlated with the position of the sentence describing that event in the paragraph. For example, the sentence appearing in the middle of the paragraph tends to have stronger relevance with visual content located around the temporal midpoint of the video. On the other hand, dense visual events mentioned in the paragraph approximately represent the global video content that unambiguously distinguishes itself from another video, which can be observed in Figure 1 (b). Therefore, inserting the query-related video into another irrelevant video automatically generates pseudo boundary labels close to the ground-truth when regarding the complete paragraph as language query for video grounding.

Motivated by the above observation, we propose a novel Siamese Grounding TRansformer (SiamGTR) for WSVPG. It jointly learns the cross-modal alignment and boundary regression via two siamese branches without generating proposals. Specifically, we propose to construct an Augmentation Branch (AB) which takes as input a pseudo video and adopts a complete paragraph as the language query to learn high-quality boundary supervision for localization. Also, an Inference Branch (IB) is designed to receive a normal video as input and is enforced to capture the order-guided cross-modal correspondence for attending over the specific video content relevant to each sentence. The two weight-sharing branches are effective to transfer complementary supervision for joint boundary prediction and feature association, which yields a weakly-supervised model with superior generalization through a concise one-stage pipeline. Extensive experiments verify the effectiveness of our model and show our method with the same or weaker supervision surpasses prior state-of-the-arts. In summary, our contributions are:

- We introduce the task of Weakly-Supervised Video Paragraph Grounding (WSVPG), which aims to train a model for localizing multiple events indicated by queries without the supervision of timestamp labels.
- We propose a novel Siamese Grounding TRansformer (SiamGTR) for concise and efficient one-stage weakly-supervised learning of video paragraph grounding. It is composed of two weight-sharing branches including an Augmentation Branch (AB) for learning boundary regression of pseudo boundaries and an Inference Branch (IB) for learning order-guided cross-modal feature alignment.
- Extensive experiments verify the efficacy of our method, and demonstrate that our framework under the same or weaker supervision outperforms state-of-the-arts.

2. Related Work

2.1. Video Sentence Grounding

Fully-Supervised Video Sentence Grounding. Plenty of approaches [1, 7, 12, 22, 29, 40, 43, 44, 47, 53, 64, 73, 74, 78, 79, 84–86, 90, 91, 93, 95–97, 100–103] have been proposed to address Fully-Supervised Video Sentence Grounding (FSVSG). In general, these works can be roughly categorized into proposal-based and proposal-free methods. Specifically, proposal-based methods [1, 22, 40, 43, 44, 73, 74, 78, 85, 86, 90, 93, 95, 101, 102] involve a proposal generation stage using sliding windows [22], anchor proposals [12, 43, 44, 47, 74, 90, 93, 95] or 2D temporal maps [64, 73, 78, 85, 101, 102], after which the generated proposals are ranked according to the query matching scores with potential post-processing like Non-Maximum Suppression (NMS). In contrast, proposal-free methods [7, 53, 79, 91, 96, 97, 100, 103] remove the dense proposal generation and score ranking process by directly regressing timestamps [7, 53, 84, 91, 103], predicting boundary distributions [96, 97, 100] or using reinforcement learning [79], which improves the computation efficiency and scenario adaptability. Following the line of proposal-free works, we propose a novel weakly-supervised regression-based framework that shows superior performance and practicability.

Weakly-Supervised Video Sentence Grounding. Weakly-Supervised Video Sentence Grounding (WSVSG) [8, 11, 26, 28, 42, 49, 52, 65, 68, 75–77, 88, 104–107] has become a popular research area because of the severe dependence of FSVSG approaches on laborious and expensive manual temporal annotations. Most of existing WSVSG methods are based on a two-stage pipeline using multiple instance learning [26, 49, 75, 88], reconstruction learning [42], or the combination of both [106, 107]. In particular, Chen *et al*. [11] have proposed a video composition strategy to generate pseudo temporal labels for WSVSG, which is the most related work to ours. However, several inherent drawbacks are involved in this approach. Firstly, individual sentences only describe local video content, thus viewing the starting/ending locations of foreground video as temporal boundaries of an individual sentence produces a weak and noisy temporal alignment, which is unsuitable for accurate

boundary supervision. Moreover, it simply adapts an existing fully-supervised proposal-based framework for weakly-supervised training, which leads to inferior generalization caused by the large train-test discrepancy. Distinct from all of the above works, we design a novel siamese framework to capture the essential characteristics of video paragraph grounding for weakly-supervised learning.

2.2. Video Paragraph Grounding

Video Paragraph Grounding (VPG) is introduced by Bao *et al*. [4], which aims to jointly localize multiple sentences of a paragraph from an untrimmed video. Shi *et al*. [62] presented an end-to-end network by re-purposing transformers into language-conditioned regressors. Jiang *et al*. [31] proposed to employ contrastive encoders for contrastive learning between video-paragraph pairs. Tan *et al*. [67] proposed a hierarchical semantic correspondence network for modeling hierarchical video-language alignment and grounding multiple levels of language queries in the video. Particularly, Jiang *et al*. [31] first explored Semi-Supervised Video Paragraph Grounding (SSVPG) to relieve the annotation burden of temporal labels. However, SSVPG is still not quite practical considering the expensive cost of temporal annotations in untrimmed videos. Besides, these semi-supervised methods still require a considerable proportion of temporal labels up to at least 10% [31] for training. To thoroughly get rid of the temporal annotations, we pioneer to explore the weakly-supervised setting in VPG.

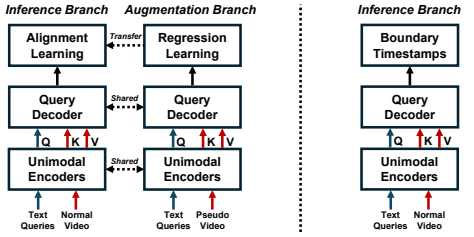
2.3. Siamese Networks

Siamese networks [6] are weight-sharing neural networks. There have been a wide range of scenarios where siamese networks are applied for achieving different purposes, such as face verification [66], image recognition [33], object tracking [5], etc. In particular, siamese networks are commonly used in contrastive self-supervised learning methods [2, 10, 13–15, 24, 25, 69, 81, 108], in which augmented views of the same or different instance are forwarded into multiple weight-sharing network copies for learning a generalizable visual representation via instance-level discrimination. In this work, we explore a new way to combine the transferability of siamese architectures and the flexibility of transformer architectures for concise and efficient weakly-supervised learning of video paragraph grounding.

3. Methodology

3.1. Overview

Task Formulation. Given an untrimmed video $\mathcal{V}$ and a paragraph query $\mathcal{P}$ consisting of N temporally ordered sentences $\{\mathcal{S}_i\}_{i=1}^{N}$, the goal of Video Paragraph Grounding (VPG) is to simultaneously localize the temporal intervals $\mathcal{T} = \{(\tau_i^{\text{st}}, \tau_i^{\text{ed}})\}_{i=1}^{N}$ of all the events described by sentences



(a) Training pipeline (b) Testing pipeline

Figure 2. Our siamese framework for joint alignment and regression.

in the paragraph, where τ_i^{st} and τ_i^{ed} are the starting and ending timestamps for the i-th sentence $\mathcal{S}_i$, respectively.

Siamese Learning. An overview of our siamese framework is shown in Figure 2. Overall, we jointly train an augmentation branch and an inference branch with shared parameters and structures for learning the complementary abilities of cross-modal feature alignment and temporal boundary regression. These two branches follow the same workflow to first encode the text queries and input video into unimodal features, after which the query features are iteratively used in the transformer decoder to extract relevant information from the video features for timestamp decoding. For testing, we only keep the pipeline of inference branch for boundary prediction. More architectural details are illustrated in Figure 3 and are elaborated in the following sections.

3.2. Feature Extraction

Video Feature Extraction. For each input video, we divide it into consecutive clips consisting of a fixed number of frames for feature extraction. Specifically, a frozen pre-trained 3D Convolutional Neural Network (3D-CNN) [70] and a linear projection layer are successively employed to obtain a 1D feature vector for each short clip, resulting in a video feature sequence $\mathcal{F}_{\text{v}} \in \mathbb{R}^{L \times D}$, where L and D are the sequence length and hidden dimension, respectively.

Text Feature Extraction. For each input paragraph consisting of N sentences, we first utilize a frozen pre-trained word embedding model to tokenize and embed the text into a sequence of word vectors. Then, a bidirectional Gated Recurrent Unit (GRU) [16] is employed on each sentence, and the last hidden states in both directions are concatenated and then projected by a linear layer to construct the sentence features $\mathcal{F}_{\text{s}} \in \mathbb{R}^{N \times D}$, where D is the hidden dimension.

3.3. Augmentation Branch

The augmentation branch aims to learn accurate boundary regression from pseudo videos with paragraph queries, which naturally transfers to the inference branch via the shared feature space established by the siamese structure.

Pseudo Data Generation. To drive end-to-end weakly-supervised regression learning, the input stream of the augmentation branch should provide reliable and direct boundary supervision. Drawing inspiration from the boundary-sensitive video pretext tasks [11, 54, 87], we propose to uti-

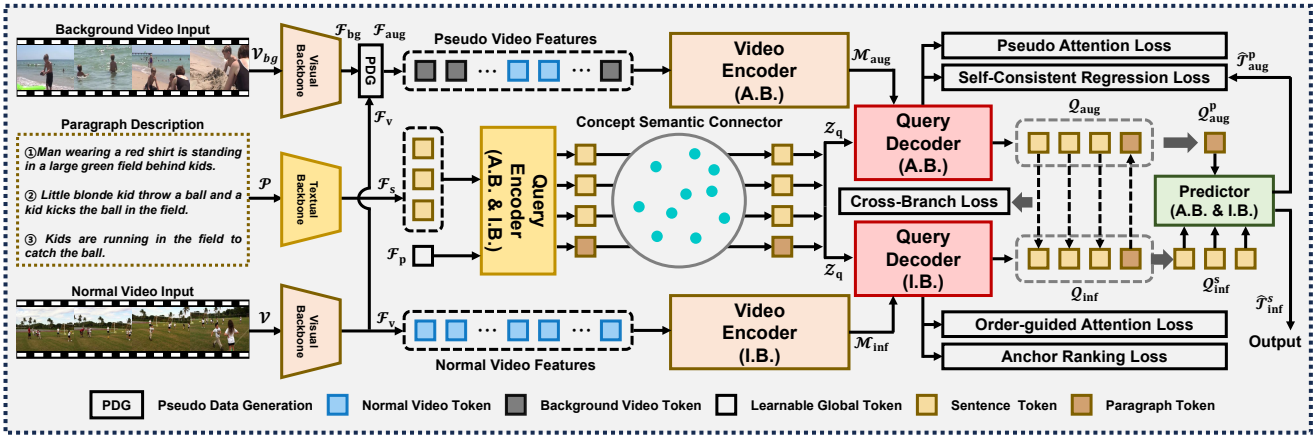


Figure 3. Illustration of the proposed Siamese Grounding TRansformer (SiamGTR) architecture. The augmentation branch (abbreviated as A.B.) takes the pseudo video features derived from randomly inserting the query-related video features into irrelevant video features. It learns to temporally regress the interval of interest from the pseudo video with the paragraph as query. The Inference Branch (abbreviated as I.B.) receives normal video features for learning the cross-modal feature alignment among multiple sentences in the video.

lize the synthesized videos paired with paragraph queries to serve as a well-suited source of surrogate boundary supervision. Specifically, for each input video $\mathcal{V}$ and its paragraph description $\mathcal{P}$, we randomly sample a background video $\mathcal{V}_{\text{bg}}$ to obtain irrelevant video content to $\mathcal{P}$. Because $\mathcal{P}$ could be treated as a unique referring expression specific to $\mathcal{V}$, the search space of background videos can be the entire training set for increasing the data diversity. Denote video features of $\mathcal{V}$ and $\mathcal{V}_{\text{bg}}$ as $\mathcal{F}_{\text{v}}$ and $\mathcal{F}_{\text{bg}}$ respectively, we construct the pseudo video features $\mathcal{F}_{\text{aug}}$ as:

$$\mathcal{F}_{\text{aug}} = \text{NRS}\left(\text{Concat}\left(\mathcal{F}_{\text{bg}}^{1:I}, \text{RRS}(\mathcal{F}_{\text{v}}), \mathcal{F}_{\text{bg}}^{I+1:}\right)\right) \quad (1)$$

where I is a randomly generated inserting index within the feature sequence $\mathcal{F}_{\text{bg}}$ and $\text{Concat}(\cdot)$ denotes the temporal concatenation. $\text{RRS}(\cdot)$ and $\text{NRS}(\cdot)$ respectively represent a Random Re-Sampling operation to stochastically re-scale the length of $\mathcal{F}_{\text{v}}$ and a Normalized Re-Sampling operation that converts the length of $\mathcal{F}_{\text{aug}}$ into a fixed number of T. Thereafter, we compute the synthesized temporal boundaries $(\tau_{\text{aug}}^{\text{st}}, \tau_{\text{aug}}^{\text{ed}})$ in the pseudo video $\mathcal{V}_{\text{aug}}$ as follows:

$$\tau_{\text{aug}}^{\text{st}} = \frac{I + \Delta I^{\text{st}}}{rL + L_{\text{bg}}}, \tau_{\text{aug}}^{\text{ed}} = \frac{I + rL - \Delta I^{\text{ed}}}{rL + L_{\text{bg}}} \quad (2)$$

where L and L_{bg} respectively indicate the length of $\mathcal{F}_{\text{v}}$ and $\mathcal{F}_{\text{bg}}$ with r being the random re-scaling factor of $\text{RRS}(\cdot)$. To alleviate potential synthesis artifacts and boundary uncertainty, we further introduce a Random Boundary Shifting (RBS) strategy that incorporates small random offsets ΔI^{st} and ΔI^{ed} into computing the pseudo labels, which is simple yet effective to boost the quality of boundary supervision.

Video Encoder. Based on the obtained pseudo video features $\mathcal{F}_{\text{aug}}$, we then encode the temporal contextual information across multiple clips by feeding $\mathcal{F}_{\text{aug}}$ to a video encoder, which follows a similar architecture of DETR encoders [9, 46, 71], i.e., each encoder layer consists of a multi-head self-attention module equipped with additive sinusoidal positional encodings and a feed-forward network. For simplicity, we omit the layer index and denote the input of each video encoder layer as $\mathcal{X}_{\text{aug}}$, then a set of Modulated Positional Encodings (MPE) are given as:

$$\mathcal{X}_{\text{pe}} = \text{MLP}(\mathcal{X}_{\text{aug}}) \odot \text{PE}(\mathcal{X}_{\text{aug}}) \quad (3)$$

where $\mathcal{X}_{\text{pe}} \in \mathbb{R}^{T \times D}$ and $\odot$ denotes element-wise multiplication. $\text{PE}(\cdot)$ is the sinusoidal function [71] with respect to temporal locations of $\mathcal{X}_{\text{aug}}$, $\text{MLP}(\cdot)$ denotes a two-layer feed-forward network computed on $\mathcal{X}_{\text{aug}}$. In each encoder layer, $\mathcal{X}_{\text{pe}}$ is only added to the input of projection layers of queries and keys in the self-attention mechanism. The encoded video features of the last video encoder layer, also called the encoder memory features $\mathcal{M}_{\text{aug}}$, are further fed to the query-guided decoder for localization prediction.

Query Encoder. Since there are strong contextual correlations among multiple sentences in the paragraph query, here we utilize a vanilla transformer encoder [71] to help reason the semantic and chronological relationships of events. To extract a global representation of the complete paragraph, we initialize an extra learnable query token $\mathcal{F}_{\text{p}}$ and integrate it with the sentence features $\mathcal{F}_{\text{s}}$ by concatenation. Thus, the input of the query encoder is constructed as follows:

$$\mathcal{X}_{\text{q}} = \left[\mathcal{F}_{\text{p}}, \mathcal{F}_{\text{s}}^{1}, \mathcal{F}_{\text{s}}^{2}, \ldots, \mathcal{F}_{\text{s}}^{i}, \ldots \mathcal{F}_{\text{s}}^{N}\right] \quad (4)$$

where $\mathcal{X}_{\text{q}} \in \mathbb{R}^{(N+1) \times D}$ and $\mathcal{F}_{\text{s}}^{i}$ is the i-th sentence feature in $\mathcal{F}_{\text{s}}$. We first normalize $\mathcal{X}_{\text{q}}$, add fixed sinusoidal positional encodings to it, and then iteratively employ self-attention modules and feed-forward networks to contextualize the global and local tokens that respectively represent the query semantics of the entire paragraph and the sentences. The output features of the last query encoder layer are denoted as $\mathcal{Z}_{\text{q}}$ and are further forwarded to the query decoder.

Conceptual Semantic Connector. Although the siamese network structure can implicitly transfer boundary knowl-

edge from the augmentation branch to the inference branch, there is still a certain semantic gap between the query representations of short sentences and long paragraphs. To narrow this gap, we develop a Conceptual Semantic Connector (CSC) module for explicit semantic guidance. Specifically, we first collect high-frequency linguistic concepts, including verbs and nouns from the training corpus, and then construct a set of dictionary features by selecting and projecting the Glove vectors [56]. The loss $\mathcal{L}_{\text{csc}}$ is computed as:

$$\mathcal{L}_{\text{csc}} = \text{BCE}\left(y^{\text{p}}_{\text{cept}}, \hat{y}^{\text{p}}_{\text{cept}}\right) + \text{BCE}\left(y^{\text{s}}_{\text{cept}}, \hat{y}^{\text{s}}_{\text{cept}}\right) \quad (5)$$

where $\text{BCE}\left(\cdot\right)$ is the binary cross-entropy loss. $y^{\text{p}}_{\text{cept}}$ and $y^{\text{s}}_{\text{cept}}$ are multi-hot labels indicating semantic concepts contained by the paragraph and sentence queries, respectively. $\hat{y}^{\text{s}}_{\text{cept}}$ and $\hat{y}^{\text{p}}_{\text{cept}}$ are the concept predictions obtained by dot-product between the conceptual dictionary features and the textual query features $\mathcal{Z}_{\text{q}}$ with a sigmoid activation.

Query Decoder. Inspired by the spiritual ideas of dynamic-anchor DETR decoders [39, 46, 51, 98], we design a novel query decoder that enables dynamic position-aware decoding of language queries. Specifically, each query decoder layer consists of a self-attention module, a cross-attention module, and a feed-forward network, with dynamically adjustable anchor boxes to indicate the query-specific location information. Initially, the input anchor boxes are set to all zeros, and an MLP is used to estimate a set of seed anchor boxes based on the cross-modal interactions between $\mathcal{M}_{\text{aug}}$ and $\mathcal{Z}_{\text{q}}$, i.e., the output features of the first decoder layer $\mathcal{Q}^{(1)}_{\text{aug}} \in \mathbb{R}^{(N+1)\times D}$ are mapped into the seed anchor boxes $\mathcal{A}^{(1)}_{\text{aug}} \in \mathbb{R}^{(N+1)\times 2}$ by the MLP. For the $(i+1)$-th decoder layer, we first convert the box coordinates of the input anchors $\mathcal{A}^{(i)}_{\text{aug}}$ into high-dimensional sinusoidal embeddings $\mathcal{F}^{(i)}_{\text{a}} \in \mathbb{R}^{(N+1)\times D}$ by a sinusoidal function [46, 71] and further obtain $\mathcal{H}^{(i)}_{\text{a}}$ by projecting $\mathcal{F}^{(i)}_{\text{a}}$ with an MLP. Afterwards, we conduct the self-attention operation and update the query features in the decoder as follows:

$$\mathcal{Q}^{(i+1)}_{\text{aug}} \leftarrow \text{Self-Attn}\begin{cases} Q = \varphi^c_q\left(\mathcal{Q}^{(i)}_{\text{aug}}\right) + \varphi^p_q\left(\mathcal{H}^{(i)}_{\text{a}}\right) \\ K = \varphi^c_k\left(\mathcal{Q}^{(i)}_{\text{aug}}\right) + \varphi^p_k\left(\mathcal{H}^{(i)}_{\text{a}}\right) \\ V = \varphi^c_v\left(\mathcal{Q}^{(i)}_{\text{aug}}\right) \end{cases} \quad (6)$$

where $\mathcal{Q}^{(i+1)}_{\text{aug}}$ is the updated query features after the self-attention operation in the $(i+1)$-th decoder layer. The above series of φ functions are used to indicate different linear projection layers for the content part or position part of the queries, keys, or values in the self-attention mechanism. Then, we conduct a cross-attention operation to extract useful cross-modal interactive information as follows:

$$\mathcal{Q}^{(i+1)}_{\text{aug}} \leftarrow \text{Cross-Attn}\begin{cases} Q = \left[\varphi^c_q\left(\mathcal{Q}^{(i+1)}_{\text{aug}}\right); \varphi^p_q\left(\mathcal{F}^{(i)}_{\text{a}}\right)\right] \\ K = \left[\varphi^c_k\left(\mathcal{M}_{\text{aug}}\right); \varphi^p_k\left(\mathcal{F}^{\mathcal{M}}_{\text{pe}}\right)\right] \\ V = \varphi^c_v\left(\mathcal{M}_{\text{aug}}\right) + \varphi^p_v\left(\mathcal{F}^{\mathcal{M}}_{\text{pe}}\right) \end{cases} \quad (7)$$

where $\mathcal{F}^{\mathcal{M}}_{\text{pe}} = \text{PE}\left(\mathcal{M}_{\text{aug}}\right)$ and the query features $\mathcal{Q}^{(i+1)}_{\text{aug}}$ is then further updated by a feed-forward network as the feature output of the $(i+1)$-th decoder layer. Then the anchor boxes are dynamically updated as $\mathcal{A}^{(i+1)}_{\text{aug}} \leftarrow \mathcal{A}^{(i)}_{\text{aug}} + \Delta\mathcal{A}^{(i)}_{\text{aug}}$, where we utilize an MLP layer to predict the relative offsets, i.e., $\Delta\mathcal{A}^{(i)}_{\text{aug}} \in \mathbb{R}^{(N+1)\times 2}$, based on the updated query features $\mathcal{Q}^{(i+1)}_{\text{aug}}$. $\mathcal{A}^{(i+1)}_{\text{aug}}$ continues to be forwarded to the next decoder layer for computing $\mathcal{F}^{(i+1)}_{\text{a}}$ and $\mathcal{H}^{(i+1)}_{\text{a}}$.

Boundary Prediction. Based on the output features and attention weights of the last query decoder layer, we simply use an MLP predictor to predict the paragraph timestamps, i.e., $\widehat{\mathcal{T}}^{\text{p}}_{\text{aug}} = (\hat{\tau}^{\text{st}}_{\text{aug}}, \hat{\tau}^{\text{ed}}_{\text{aug}})$. Similarly, the sentence timestamps $\widehat{\mathcal{T}}^{\text{s}}_{\text{inf}} = \left\{(\hat{\tau}^{\text{st}}_{\text{j}}, \hat{\tau}^{\text{ed}}_{\text{j}})\right\}^{N}_{j=1}$ can also be obtained by feeding last-layer output from the inference branch to the same MLP.

Self-Consistent Boundary Regression. We improve the attention-agnostic regression loss $\mathcal{L}_{\text{reg}}$ to make it aware of the model's self-consistent scores, where the main idea is to selectively optimize the regression loss of self-consistent samples for better weakly-supervised regression learning. Self-consistent samples have high attention weights over the pseudo ground-truth intervals, which are more suitable for learning less noisy supervision for accurate boundary prediction. Specifically, the self-consistent boundary regression loss $\mathcal{L}_{\text{screg}}$ is defined as:

$$\mathcal{L}_{\text{screg}} = \begin{cases} \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{GIoU}}, & \text{if} \quad s_{\text{att}} > \beta \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{GIoU}}$ respectively represent L1 and Generalized Intersection over Union (GIoU) [58] loss computed between $(\hat{\tau}^{\text{st}}_{\text{aug}}, \hat{\tau}^{\text{ed}}_{\text{aug}})$ and $(\tau^{\text{st}}_{\text{aug}}, \tau^{\text{ed}}_{\text{aug}})$. β is set to 0.5 and s_{att} is the attention sum over the pseudo ground-truth interval.

3.4. Inference Branch

In our siamese framework, the inference branch shares the same parameter weights and network structure with the augmentation branch, i.e., a video encoder, a query encoder, and a query decoder. The only difference lies in the input streams and objectives, i.e., the inference branch receives a normal video during training to learn in-domain cross-modal correspondence that cannot be acquired through the pseudo video stream, which significantly improves the generalization ability of the model. Specifically, it receives the encoded normal video features $\mathcal{M}_{\text{inf}}$ and the encoded text query features $\mathcal{Z}_{\text{q}}$ as the decoder input and generates the hidden query features $\mathcal{Q}_{\text{inf}}$ for boundary prediction.

Order-guided Attention Loss. The chronological prior given by the sentence order provides explicit guidance for learning cross-modal alignment between the video and language features during decoding. To learn the order-guided cross-modal correspondence, we constrain cross-modal attention weights in the decoder as follows:

$$\mathcal{L}_{\text{oga}} = max\left(0, \Delta mT + \sum_{t=1}^{T} t\alpha_{\text{s}}^{j}(t) - \sum_{t=1}^{T} t\alpha_{\text{s}}^{j+1}(t)\right) \quad (9)$$

where $\alpha_{\text{s}}^{j}(t)$ and $\alpha_{\text{s}}^{j+1}(t)$ are attention weights over video features for the j-th and $(j+1)$-th sentence from the last decoder layer, respectively. Δm is the minimal distance between attention centroids. This loss is mainly contributed by the inference branch with $\Delta m = \frac{1}{2N}$ and partly contributed by the augmentation branch with $\Delta m = \frac{1}{4N}$.

Auxiliary Losses. To exploit more guidance for weakly-supervised representation learning, we employ three auxiliary losses including a cross-branch loss $\mathcal{L}_{\text{cb}}$, an anchor ranking loss $\mathcal{L}_{\text{ar}}$ and a pseudo attention loss $\mathcal{L}_{\text{pa}}$. Specifically, $\mathcal{L}_{\text{cb}}$ utilizes the semantic consistency constraint of output features across siamese branches, which is calculated analogous to MoCo [25]. $\mathcal{L}_{\text{ar}}$ is used for inducing the decoder to learn a set of order-preserving anchor boxes, which is computed on the anchor boxes like in the equation (9). $\mathcal{L}_{\text{pa}}$ makes use of attention supervision between the pseudo video and language features from the augmentation branch, which follows the calculation proposed in [62, 91].

3.5. Training and Inference

Weakly-Supervised Learning. The weakly-supervised loss of our proposed framework can be formulated as a weighted sum of the two losses from the siamese branches as $\mathcal{L}_{\text{WS}} = \lambda_{\text{screg}}\mathcal{L}_{\text{screg}} + \lambda_{\text{oga}}\mathcal{L}_{\text{oga}}$, where λ_{screg} and λ_{oga} are scalar weights to balance the contributions of the two different losses. The overall loss for the weakly-supervised model is defined as $\mathcal{L} = \mathcal{L}_{\text{WS}} + \lambda_{\text{csc}}\mathcal{L}_{\text{csc}} + \mathcal{L}_{\text{aux}}$.

Semi-Supervised Learning. Although our framework is initially designed for weakly-supervised learning, it can be easily adapted for end-to-end semi-supervised learning. Specifically, in addition to $\mathcal{L}_{\text{WS}}$ which is calculated on all training samples, we only need to employ an extra fully-supervised loss $\mathcal{L}_{\text{FS}}$ on those fully-annotated samples without changing any part of the network structure. The semi-supervised loss $\mathcal{L}_{\text{SS}}$ is defined as $\mathcal{L}_{\text{SS}} = \mathcal{L}_{\text{WS}} + \mathcal{L}_{\text{FS}}$, where $\mathcal{L}_{\text{FS}}$ consists of a regression loss and an attention loss and is calculated on the labeled samples in the inference branch.

Model Inference. As mentioned, the augmentation branch and conceptual semantic connector are discarded, while other inference branch modules are preserved for testing.

4. Experiment

4.1. Datasets and Metrics

ActivityNet-Captions. ActivityNet-Captions [34] dataset is a large-scale dataset with diverse open-domain content sourced from ActivityNet dataset [34]. There are 14,926 videos and 19,811 localized video-paragraph pairs in total. Each video lasts for 117.60 seconds and each paragraph consists of 3.63 sentences on average. The entire dataset is divided into train/val_1/val_2 sets containing 10,009/4,917/4,885 video-paragraph pairs, respectively. We follow prior works [4, 31] to use val_2 set for testing.

Charades-CD-OOD. Charades-STA dataset [22] is built from the Charades dataset [63] with indoor activities. Following the previous work [31], we adopt a reorganized version of Charades-STA named Charades-CD-OOD proposed in [92]. It is divided into train/val/test_ood sets consisting of 4,564/333/1,440 video-paragraph pairs, respectively. Specifically, the average video duration is 30.78 seconds and the average paragraph length is 2.41 sentences.

TACoS. TACoS dataset [57] is constructed from the MPII corpus [61] tailored for cooking activities and kitchen scenarios. There are 127 videos in total with each video paired with multiple paragraphs at different granularities. Concretely, there are 1,107, 418, and 380 video-paragraph pairs for training, validation, and testing, respectively. The average video length and number of sentences in the paragraph are 4.79 minutes and 8.75 in this dataset, respectively.

Evaluation Metrics. Following previous works [4, 31], we adopt mean Intersection over Union (i.e., mIoU) and recall under IoU threshold of m (i.e., R@m) as our evaluation metrics. The metrics are averaged over all sentences and m is set to be $\{0.3, 0.5, 0.7\}$ for ActivityNet-Captions and Charades-CD-OOD, and $\{0.1, 0.3, 0.5\}$ for TACoS.

4.2. Implementation Details

For fair comparison with existing works [4, 62], we adopt the same C3D network [70] and Glove model [56] as feature extractors. The number of sampled video clips T is set to be 256, 128, and 512 for ActivityNet-Captions, Charades-CD-OOD, and TACoS datasets, respectively. We train the model using Adam [32] optimizer with a fixed learning rate of 0.0001 and a batch size of 32, 32, and 16 for ActivityNet-Captions, Charades-CD-OOD and TACoS, respectively. We select top-100 high-frequency concepts for each dataset, and the loss weights $\{\lambda_{\text{screg}}, \lambda_{\text{oga}}, \lambda_{\text{csc}}\}$ are set to $\{2, 1, 10\}$. The number of encoder and decoder layers is set to be 3, and the hidden size D is 256 in all settings.

4.3. Comparison with State-of-the-arts

We compare the proposed SiamGTR with existing state-of-the-art methods for VPG to demonstrate the superiority of our framework. Specifically, 3D-TPN [4, 101], DepNet [4], PRVG [62], SVPTR [31] and HSCNet [67] are fully-supervised approaches requiring temporal annotations for the entire dataset. Besides, the semi-supervised setting has been studied in [31] with several methods developed. For fair comparison with our method, we regard the reconstruction learning method WSSL [18] as one baseline and further develop a more competitive model called Weakly-Supervised Temporal Paragraph Network (WSTPN) by incorporating Beam Search [20] into WSTAN [75]. Specifically, WSTPN utilizes a complete paragraph for multiple

Table 1. Comparison on ActivityNet-Captions dataset.

Method	Setting	R@0.3	R@0.5	R@0.7	mIoU
3D-TPN [101]	FS	67.56	51.49	30.92	-
DepNet [4]	FS	72.81	55.91	33.46	-
PRVG [62]	FS	78.27	61.15	37.83	55.62
SVPTR [31]	FS	78.07	61.70	38.36	55.91
HSCNet [67]	FS	81.89	66.57	44.03	59.71
DepNet [4]	SS	61.46	45.14	26.78	44.11
VPTR [31]	SS	72.80	53.14	29.07	50.08
SVPTR [31]	SS	73.39	56.72	32.78	51.98
SiamGTR (Ours)	SS	**78.75**	**59.11**	**34.12**	**54.57**
WSSL [18]	WS	41.98	23.34	-	28.33
WSTPN [75]	WS	57.74	33.02	13.62	38.54
SiamGTR (Ours)	WS	**75.43**	**57.23**	**30.56**	**52.32**

Table 2. Comparison on Charades-CD-OOD dataset.

Method	Setting	R@0.3	R@0.5	R@0.7	mIoU
DepNet [4]	FS	45.61	27.59	10.69	29.30
STLG [48]	FS	48.30	30.39	9.79	-
SVPTR [31]	FS	55.14	32.44	15.53	36.01
DepNet [4]	SS	43.03	25.07	10.14	28.09
STLG [48]	SS	46.15	29.43	9.38	-
VPTR [31]	SS	45.13	24.98	10.22	28.92
SVPTR [31]	SS	50.31	28.50	12.27	32.13
SiamGTR (Ours)	SS	**59.07**	**35.47**	**14.95**	**38.87**
WSSL [18]	WS	35.86	23.67	8.27	-
WSTPN [75]	WS	48.61	29.27	10.79	33.49
SiamGTR (Ours)	WS	**57.33**	**33.87**	**12.31**	**37.21**

instance learning and searches the best sequence of timestamps with the highest overall confidence while maintaining a consistent temporal order with the input sentences.

Comprehensive results over three different datasets are shown in Table 1, Table 2, and Table 3, where FS/SS/WS are used to indicate fully/semi/weakly-supervised settings of video paragraph grounding, respectively. First of all, our SiamGTR remarkably surpasses all the other methods under the same supervision in all metrics over the three datasets. Concretely, our framework outperforms WSTPN by 13.78%, 3.72%, and 11.16% in mIoU on ActivityNet-Captions, Charades-CD-OOD and TACoS datasets, respectively. Compared to semi-supervised methods using a considerable number of temporal labels, our weakly-supervised method is also able to achieve comparable or even better results, which demonstrates the effectiveness of our framework in efficient weakly-supervised learning. Furthermore, our framework is flexible and can be easily adapted to semi-supervised learning for further gains. As shown, our semi-supervised model outstrips all semi-supervised state-of-the-arts by a large margin, and it performs better or on par with the fully-supervised SVPTR on all three datasets.

4.4. Ablation Study

We conduct ablation studies to investigate the contributions of different components on ActivityNet-Captions dataset.

Effectiveness of data augmentation. To evaluate the influences of the random boundary shifting and random resampling operations for data augmentation, we remove one or both of them from the training pipeline and the results are shown in Table 4 (a) $\sim$ (d). The model performance clearly degrades after the removal, which demonstrates the necessity of increasing sample diversity and alleviating overfitting for weakly-supervised cross-modal regression learning.

Table 3. Comparison on TACoS dataset.

Method	Setting	R@0.1	R@0.3	R@0.5	mIoU
3D-TPN [101]	FS	55.05	40.31	26.54	-
DepNet [4]	FS	56.10	41.34	27.16	-
PRVG [62]	FS	61.64	45.40	26.37	29.18
SVPTR [31]	FS	67.91	47.89	28.22	31.42
HSCNet [67]	FS	76.28	59.74	42.00	40.61
DepNet [4]	SS	40.27	26.95	16.54	18.68
VPTR [31]	SS	61.31	40.59	21.39	26.59
SVPTR [31]	SS	63.06	40.19	20.05	26.10
SiamGTR (Ours)	SS	**67.30**	**49.35**	**31.69**	**32.81**
WSTPN [75]	WS	28.59	10.04	4.76	9.32
SiamGTR (Ours)	WS	**61.51**	**26.22**	**10.53**	**20.48**

Table 4. Ablation studies on component designs of our framework. Experimental results are marked from ID (a) $\sim$ (l). RBS and RRS respectively denote the random boundary shifting and random resampling operations for pseudo data generation. MPE, CSC and DAB stand for the modulated positional encodings, conceptual semantic connector, and dynamic anchor boxes, respectively.

ID	RBS	RRS	MPE	CSC	DAB	R@0.5	mIoU
(a)			✓	✓	✓	47.10	46.04
(b)	✓		✓	✓	✓	48.19	46.96
(c)		✓	✓	✓	✓	54.24	50.94
(d)	✓	✓	✓	✓	✓	57.23	52.32
(e)	✓	✓				45.25	44.46
(f)	✓	✓	✓			46.47	45.47
(g)	✓	✓		✓		46.96	45.23
(h)	✓	✓			✓	51.34	48.52
(i)	✓	✓	✓	✓		52.12	48.75
(j)	✓	✓	✓		✓	53.57	50.06
(k)	✓	✓		✓	✓	54.09	50.84
(l)	✓	✓	✓	✓	✓	57.23	52.32

Ablation on module designs. As shown in Table 4 (e) $\sim$ (l), we conduct detailed ablation studies on module designs to validate the rationality of the proposed model. As observed, the designs of modulated positional encodings and dynamic anchor boxes in the encoder and decoder are consistently beneficial to improving the model capacity for better performance. The conceptual semantic connector that bridges two types of queries is also effective, and it further boosts the performance by 2.26% in mIoU (50.06% vs. 52.32%) even though the network has been equipped with strong encoders and decoders. We notice the dynamic anchors for decoding are the most crucial component given the sharpest performance drop of 3.57% with its removal, which indicates the importance to explicitly represent the intermediate location information for video grounding.

Table 5. Ablation studies on different weakly-supervised losses.

ID	$\mathcal{L}_{\text{screg}}$	$\mathcal{L}_{\text{oga}}$	R@0.3	R@0.5	R@0.7	mIoU
(a)			45.85	29.82	10.93	30.97
(b)	✓		46.90	29.38	12.06	31.82
(c)		✓	73.58	54.58	28.82	50.62
(d)	✓	✓	75.43	57.23	30.56	52.32

Table 6. Impact of different auxiliary losses.

Method	R@0.3	R@0.5	R@0.7	mIoU
w/o $\mathcal{L}_{\text{cb}}$	71.49	49.78	24.64	48.07
w/o $\mathcal{L}_{\text{ar}}$	72.96	50.72	26.42	49.14
w/o $\mathcal{L}_{\text{pa}}$	74.34	56.01	30.29	51.76
Full Model	75.43	57.23	30.56	52.32

Table 7. Evaluation on different types of paragraph representation.

Method	R@0.3	R@0.5	R@0.7	mIoU
Max-Pooling	72.61	52.61	27.03	49.55
Mean-Pooling	74.31	53.13	28.64	50.57
Word Concat	73.79	52.27	27.30	49.96
Learnable	75.43	57.23	30.56	52.32

Analysis on weakly-supervised losses. Ablation studies on contributions of two weakly-supervised losses $\mathcal{L}_{\text{screg}}$ and $\mathcal{L}_{\text{oga}}$ are shown in Table 5. In experiment (a) and (c), we remove $\mathcal{L}_{\text{screg}}$ but use a plain regression loss $\mathcal{L}_{\text{reg}}$ for ablation analysis. Firstly, simply employing regressive supervision attains inferior performance because the feature-level cross-modal correspondence can hardly be learned with coordinate-level supervision. Furthermore, we find $\mathcal{L}_{\text{oga}}$ is critical for learning precise temporal localization since it explicitly guides the model to align video features and language features that are highly likely to be correlated. Besides, we observe that using $\mathcal{L}_{\text{screg}}$ for selecting high-quality regression samples always brings gains to the performance.

Impact of auxiliary losses. We investigate the influences of auxiliary losses on the model performance, which involve $\mathcal{L}_{\text{cb}}$ for exploiting the cross-branch knowledge, $\mathcal{L}_{\text{ar}}$ as guidance to learn a set of order-preserving anchor boxes and $\mathcal{L}_{\text{pa}}$ to make use of the feature alignment supervision from the augmentation branch. As presented in Table 6, all three auxiliary losses bring positive impacts, with $\mathcal{L}_{\text{cb}}$ being the most effective to improve the mIoU metric by 4.25%. The reason might lie in the rich complementary knowledge and consistency supervision across the siamese branches.

Impact of the paragraph representation. The comparison of different types of paragraph representation is shown in Table 7. Four different schemes are included, i.e., mean-pooling or max-pooling the sentence features, embedding the sequence of all word tokens in the paragraph, and using a learnable query token to extract global information. It is clear that adaptively learning a paragraph representation with our method achieves the best performance and the max-pooling scheme performs the worst because it loses too much detailed information. The mean-pooling scheme performs slightly better than the word-concat scheme, which may attribute to the advantage of late fusion at feature level.

Figure 4. Visualization of prediction results from different models.

4.5. Visualization

In Figure 4, we intuitively visualize the predicted timestamps from WSTPN and our proposed model trained under the weakly-supervised or semi-supervised setting. Overall, the predicted timestamps given by WSTPN coarsely capture the sentence order relations but with inaccurate boundaries. In contrast, our weakly-supervised model achieves much better results. It is notable that our semi-supervised model generates more fine-grained boundaries, which shows the advantage of our siamese framework in jointly leveraging fewer and weaker labels for efficient learning.

5. Conclusion

In this work, we explore the weakly-supervised setting in video paragraph grounding (i.e., WSVPG) to eliminate the dependence of the temporal annotations. To achieve this goal, we propose a novel siamese learning framework to jointly learn the cross-modal feature alignment and temporal coordinate regression without ground-truth supervision. Specifically, we design a novel Siamese Grounding TRansformer (SiamGTR) consisting of an augmentation branch and an inference branch. The augmentation branch utilizes the boundary supervision provided by temporally regressing a complete paragraph in a pseudo video, and the inference branch learns the order-guided cross-modal correspondence of multiple sentences in a normal video. Extensive experiments verify the effectiveness of our framework.

Acknowledgements. This work was supported partially by the NSFC (U21A20471, U22A2095, 62076260, 61772570), Guangdong Natural Science Funds Project (2020B1515120085, 2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B1515020009), and Guangzhou Science and Technology Plan Project (202201011134).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 1, 2

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 3

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1

[4] Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *AAAI*, 2021. 1, 3, 6, 7

[5] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV Workshops*, 2016. 3

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *NeurIPS*, 1993. 3

[7] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *EMNLP*, 2021. 2

[8] Meng Cao, Fangyun Wei, Can Xu, Xiubo Geng, Long Chen, Can Zhang, Yuexian Zou, Tao Shen, and Daxin Jiang. Iterative proposal refinement for weakly-supervised video grounding. In *CVPR*, 2023. 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[11] Jiaming Chen, Weixin Luo, Wei Zhang, and Lin Ma. Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. In *AAAI*, 2022. 2, 3

[12] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, 2020. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[14] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[15] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[16] Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014. 3

[17] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *ACMMM*, 2022. 1

[18] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *NeurIPS*, 2018. 6, 7

[19] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, 2019. 1

[20] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *ACL*, 2017. 6

[21] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1

[22] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 2, 6

[23] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 1

[24] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. *arXiv preprint arXiv:2305.14344*, 2023. 3

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 6

[26] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, 2021. 2

[27] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, 2020. 1

[28] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *CVPR*, 2023. 2

[29] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, 2023. 2

[30] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 1

[31] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *CVPR*, 2022. 3, 6, 7

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[33] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML Workshops*, 2015. 3

[34] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 6

[35] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *CVPR*, 2022. 1

[36] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, 2020. 1

[37] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, 2017. 1

[38] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 1

[39] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 5

[40] Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, and Yuexian Zou. G2l: Semantically aligned and uniform video grounding via geodesic and game theory. In *ICCV*, 2023. 2

[41] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, 2022. 1

[42] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, 2020. 2

[43] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *ACMMM*, 2022. 2

[44] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, 2022. 2

[45] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *ICCV*, 2023. 1

[46] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *ICLR*, 2021. 4, 5

[47] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. Debug: A dense bottom-up grounding approach for natural language video localization. In *EMNLP*, 2019. 2

[48] Fan Luo, Shaoxiang Chen, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Self-supervised learning for semi-supervised temporal language grounding. *TMM*, 2022. 7

[49] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *ECCV*, 2020. 2

[50] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *CVPR*, 2017. 1

[51] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, 2021. 5

[52] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *CVPR*, 2019. 2

[53] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 2

[54] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *ACCV*, 2022. 3

[55] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. In *NeurIPS*, 2021. 1

[56] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 5, 6

[57] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *TACL*, 2013. 6

[58] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[59] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *CVPR*, 2019. 1

[60] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, 2018. 1

[61] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *ECCV*, 2012. 6

[62] Fengyuan Shi, Limin Wang, and Weilin Huang. End-to-end dense video grounding via parallel regression. *arXiv preprint arXiv:2109.11265*, 2021. 3, 6, 7

[63] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 6

[64] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *CVPR*, 2021. 2

[65] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 2

[66] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 3

[67] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Wei-Shi Zheng, and Jianhuang Lai. Hierarchical semantic correspondence networks for video paragraph grounding. In *CVPR*, 2023. 3, 6, 7

[68] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *WACV*, 2021. 2

[69] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *CVPR*, 2023. 3

[70] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 3, 6

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4, 5

[72] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. In *AAAI*, 2021. 1

[73] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, 2021. 2

[74] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, 2020. 2

[75] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *TMM*, 2022. 2, 6, 7

[76] Yunxiao Wang, Meng Liu, Yinwei Wei, Zhiyong Cheng, Yinglong Wang, and Liqiang Nie. Siamese alignment network for weakly supervised video moment retrieval. *TMM*, 2023.

[77] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACMMM*, 2021. 2

[78] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, 2022. 2

[79] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, 2020. 2

[80] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 2023. 1

[81] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[82] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 1

[83] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. Convolutional hierarchical attention network for query-focused video summarization. In *AAAI*, 2020. 1

[84] Shaoning Xiao, Long Chen, Jian Shao, Yueting Zhuang, and Jun Xiao. Natural language video localization with learnable moment proposals. In *EMNLP*, 2021. 2

[85] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI*, 2021. 2

[86] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019. 2

[87] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martinez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *ICCV*, 2021. 3

[88] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *TIP*, 2021. 2

[89] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019. 1

[90] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, 2019. 2

[91] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019. 2, 6

[92] Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *ACMMM Workshops*, 2021. 6

[93] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, 2020. 2

[94] Chenchi Zhang, Wenbo Ma, Jun Xiao, Hanwang Zhang, Jian Shao, Yueting Zhuang, and Long Chen. Vl-nms: Breaking proposal bottlenecks in two-stage visual-language matching. *TOMM*, 2023. 1

[95] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019. 2

[96] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, 2020. 2

[97] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*, 2021. 2

[98] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022. 5

[99] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1

[100] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*, 2021. 2

[101] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020. 2, 6, 7

[102] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *TPAMI*, 2021. 2

[103] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. Text-visual prompting for efficient 2d temporal video grounding. In *CVPR*, 2023. 2

[104] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACMMM*, 2020. 2

[105] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *NeurIPS*, 2020.

[106] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, 2022. 2

[107] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*, 2022. 2

[108] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2021. 3