# HiSA: Hierarchically Semantic Associating for Video Temporal Grounding

Zhe Xu, Da Chen, Kun Wei, Cheng Deng, *Senior Member, IEEE*, and Hui Xue

***Abstract*— Video Temporal Grounding (VTG) aims to locate the time interval in a video that is semantically relevant to a language query. Existing VTG methods interact the query with entangled video features and treat the instances in a dataset independently. However, *intra-video entanglement* and *inter-video connection* are rarely considered in these methods, leading to mismatches between the video and language. To this end, we propose a novel method, dubbed Hierarchically Semantic Associating (HiSA), which aims to precisely align the video with language and obtain discriminative representation for further location regression. Specifically, the action factors and background factors are disentangled from adjacent video segments, enforcing precise multimodal interaction and alleviating the intra-video entanglement. In addition, cross-guided contrast is elaborately framed to capture the inter-video connection, which benefits the multimodal understanding to locate the time interval. Extensive experiments on three benchmark datasets demonstrate that our approach significantly outperforms the state-of-the-art methods. The project page is available at: https://github.com/zhexu1997/HiSA.**

***Index Terms*— Video temporal grounding, feature disentanglement, cross-guided contrast.**

## I. Introduction

With the rapid growth of video data, video understanding has attracted extensive research interest. Traditional video understanding tasks, *e.g.*, video classification [1]–[5], object tracking [6]–[8], and action localization [9]–[13] are limited in the video modality and pre-defined action categories. In contrast, multi-modal video understanding tasks [14]–[21] involve video and other modalities such as text and audio, which are more challenging than traditional tasks. Due to the promising applications, they have been increasingly studied in the last few years. In this paper, we focus on a typical multi-modal task called Video Temporal Grounding (VTG), which aims to locate the clip in an untrimmed video that is semantically relevant to a language query.

Manuscript received 8 January 2022; revised 27 May 2022; accepted 28 June 2022. Date of publication 1 August 2022; date of current version 4 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361; in part by the Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03; and in part by the Fundamental Research Funds for the Central Universities under Grant ZDRC2102. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Chuang Gan. *(Corresponding author: Cheng Deng.)*

Zhe Xu, Kun Wei, and Cheng Deng are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: zhexu@stu.xidian.edu.cn; weikunsk@gmail.com; chdeng.xd@gmail.com).

Da Chen and Hui Xue are with Alibaba Group, Hangzhou 311121, China (e-mail: chen.cd@alibaba-inc.com; hui.xueh@alibaba-inc.com).

Digital Object Identifier 10.1109/TIP.2022.3191841

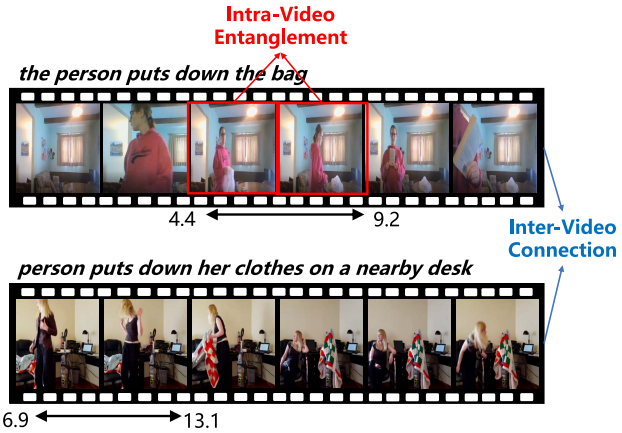



Fig. 1. Illustrations of the VTG task and two problems rarely considered by existing approaches. (1) intra-video entanglement: the action is entangled with the background within a video, making it challenging for a model to distinguish the differences between adjacent segments and precisely locate the time interval. (2) inter-video connection: different query-video pairs in a dataset are likely to be closely connected with each other, providing a model various information to facilitate a better understanding of vision and language.

Existing VTG methods can be roughly divided into three categories: (1) proposal-based methods [22], [23], (2) proposal-free methods [24]–[26], and (3) reinforcement learning-based methods [27], [28]. Proposal-based methods first generate a series of candidate video clips as proposals, then output the time interval of the proposal that is most relevant to the given language query. Since these methods need to calculate the similarities between all proposals and language queries, the computational costs become extremely high when tackling large-scale video data. Proposal-free methods treat VTG as a regression problem and directly predict the time interval leveraging the video and language information. Due to its computational efficiency, most recent VTG methods adopt the proposal-free approach. In addition, reinforcement learning-based methods formulate VTG as a sequential decision making problem, where an agent is trained to progressively regulate the temporal grounding boundaries.

For all VTG methods mentioned above, the performance heavily relies on the quality of the learned features. However, there are two problems scarcely considered by existing works: (1) intra-video entanglement, and (2) inter-video connection, which hinders VTG models from obtaining discriminative multimodal features. As the examples illustrated in Figure 1, given two language queries (*the person puts down the bag* and *person puts down her clothes on a nearby desk*) and the

untrimmed videos, intra-video entanglement naturally exists, *i.e.*, the action is entangled with the background in videos. Adjacent video segments within a video are similar in appearance, making it challenging for a model to distinguish the differences between them and precisely locate the relevant time interval. In addition, different query-video pairs in a dataset are likely to be closely connected with each other, providing the model various information to facilitate a better understanding of vision and language. Existing methods, however, treat them independently, which hinders the accurate language grounding. Thus, a simple yet effective solution to improve the grounding performance would be to simultaneously consider the intra-video entanglement and inter-video connection to achieve a precise multimodal alignment.

In this paper, we present Hierarchically Semantic Associating (HiSA) for VTG, where feature disentanglement and cross-guided contrast are performed to deal with intra-video entanglement and inter-video connection, respectively. Concretely, the action and background factors are first disentangled from video segments based on adjacent features. Subsequently, a transformer-based fusion network is presented to fuse the multimodal features in a common space. In addition, query-guided video contrast and video-guided query contrast are leveraged to establish the inter-video connection, where different cross-modal negative samples are weighted by the similarities among corresponding paired samples from the same modality.

Our key contributions can be summarized as follows:

1) We propose Hierarchically Semantic Associating (HiSA) framework for VTG, which focuses on the intra-video entanglement and inter-video connection to better align video with language.
2) To the best of our knowledge, this is the first work to apply feature disentanglement for VTG. Action and background factors in the video segments are disentangled, which alleviates the intra-video entanglement and benefits subsequent interaction of video and language.
3) We present cross-guided contrast to establish the inter-video connection, obtaining discriminative video and query representations for location regression.
4) Extensive experiments on three benchmark datasets demonstrate that our method can achieve the state-of-the-art performance for the VTG task.

## II. Related Works

### A. Video Temporal Grounding

VTG methods can be roughly divided into three categories: (1) proposal-based methods, (2) proposal-free methods, and (3) reinforcement learning-based methods. Proposal-based methods [22], [23] follow the two-stage "proposal+matching" paradigm, where a set of candidate proposals are generated for a given video and then the most relevant time interval is selected. CTRL [22] uses a dense sliding window to produce activity proposals and proposes a Cross-modal Temporal Regression Localizer to obtain alignment scores and location regression results for candidate proposals. SAP [23] proposes the Semantic Activity Proposal framework to integrate semantic information into the proposal generation process. 2D-TAN [29] proposes to model the temporal relations between video candidates by a two-dimensional map, where one dimension indicates the starting time of a moment and the other indicates the end time.

One of the major limitations of proposal-based methods is that it is computationally expensive to compare all the proposals with the language query. Therefore, numerous proposal-free methods [24], [25], [30], [31] have been proposed to treat VTG as a regression problem and directly predict the time interval. ABLR [24] introduces a co-attention mechanism for VTG and proposes Attention Based Location Regression to regress the temporal coordinates of the language query from attention weights or attended features. LGI [25] extracts multiple semantic phrases from a language query and presents Local Global Interactions to fuse multimodal information. ACRM [32] proposes an Attentive Cross-modal Relevance Matching model, which introduces an attention mechanism to model the fine-grained interactions between the video frames and the query words.

Reinforcement learning-based methods formulate VTG as a sequential decision making problem. RWM [28] views VTG as controlling an agent to read the description, to watch the video as well as the current localization, and then to iteratively move the temporal grounding boundaries to find the best matching clip. SM-RL [33] proposes a recurrent neural network based reinforcement learning model that selectively observes a sequence of frames and associates the given sentence with video content in a matching-based manner. TSP-PRL [34] presents a Tree-Structured Policy based Progressive Reinforcement Learning framework to sequentially regulate the temporal boundary by an iterative refinement process.

Despite great progress, none of the existing methods considers the intra-video entanglement and inter-video connection, which is essential to better align the video and language query.

### B. Disentangled Representation Learning

Disentangled representation learning [35], [36] aims to model the factors of data variations, which can be applied to various tasks such as image-to-image translation [37]–[40], domain adaptation [41], and face recognition [42]. InfoGAN [43] achieves disentanglement by maximizing the mutual information between latent variables and input data. Other works [44], [45] employ VAEs [46] for disentanglement. Recently, S3VAE [47] is proposed to disentangle time-invariant and time-varying representations for sequential data. However, S3VAE disentangles the instances in a sequence independently and thus requires auxiliary supervision, *e.g.*, optical flow, to guarantee disentanglement. In addition, some self-supervised video representation learning works [48], [49] also focus on learning action and background features in a disentanglement way, which are achieved by frame difference or relative speed perception.

Different from the methods discussed above, we disentangle video segments into action factors and background factors by employing our feature disentanglement network, where the inductive bias is introduced by adjacent features to guarantee the disentanglement.

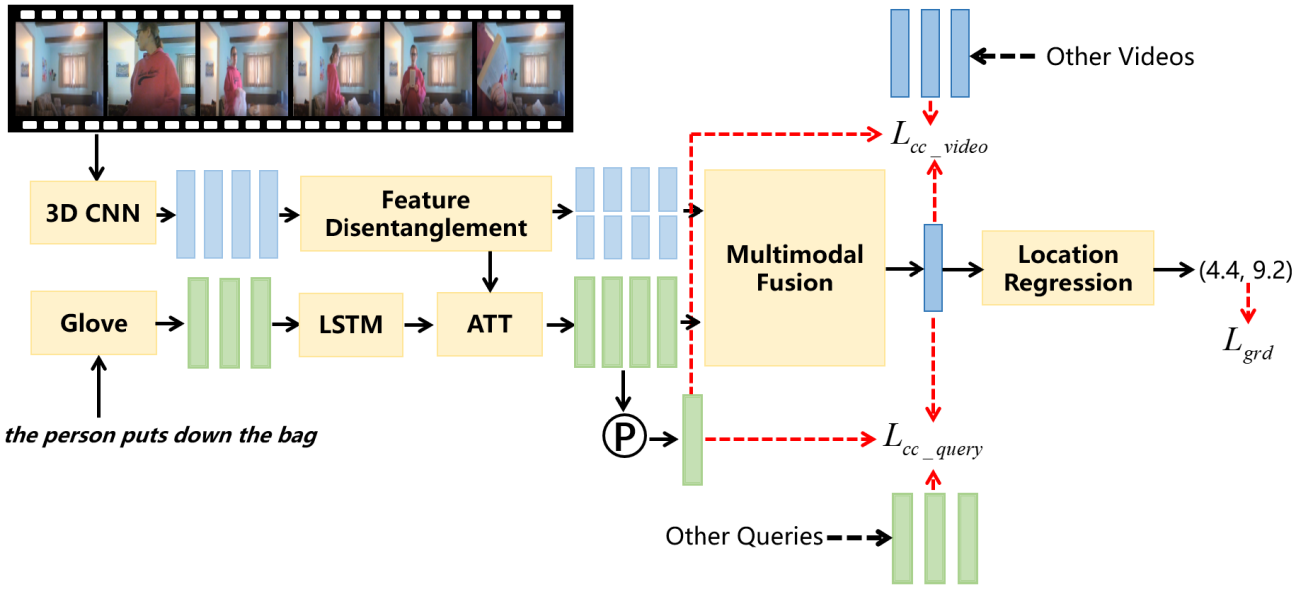



Fig. 2. The framework of our proposed HiSA for VTG. Given video segment features and word-level query features extracted from pre-trained models, we disentangle the video segment into action factors as well as background factors by feature disentanglement network and obtain segment-aware query feature by LSTM and an attention network. The the transformer-based multimodal fusion network is then employed to generate fused video features, where the time interval is regressed. Cross-guided contrast is performed to better align the video with the language query.

### C. Contrastive Learning

Contrastive learning [50]–[52] has recently achieved success in the self-supervised representation learning context. MoCo [51] builds dynamic dictionaries for self-supervised visual representation learning. SimCLR [52] systematically studies the components of the framework, including the composition of data augmentations, non-linear projection head, and batch size. In addition to self-supervised representation learning, SupCon [53] shows that contrastive learning can also boost the supervised classification performance. IVG-DCL [26] introduces a dual contrastive learning approach for VTG and is the work most closely related to our method. However, IVG-DCL only performs contrastive learning within a video and treats negative samples equally.

In contrast, our proposed cross-guided contrast for VTG is superior in three aspects: (1) We generate negative samples from instances in the minibatch, whose quantity is not limited within a single video. (2) We perform contrastive learning for both video and query representation learning. (3) We attach different weights for negative samples.

## III. Proposed Method

Given an untrimmed video $V$ and a language query $Q$, $I = (t^s, t^e)$ represents the time interval in the video corresponding to the query, where $t^s$ is the start time and $t^e$ is the end time. VTG aims to learn a model that can predict the time interval $I$ based on the untrimmed video $V$ and language query $Q$:

$$\theta^* = \arg\max_{\theta} \mathbb{E}[\log p_\theta(I|V, Q)], \tag{1}$$

where $\log p_\theta(I|V, Q)$ is the likelihood function and $\theta$ is the model parameter.

Figure 2 presents the framework of our proposed method. Video segment features and text word features are first extracted from pre-trained video and language models, respectively. We then disentangle the video segments into action factors and background factors, and obtain segment-aware query representation by LSTM and an attention network. Next, the features from these two modalities are fed into the fusion network to generate the fused video segment feature, where the time interval is regressed. In addition, cross-guided contrast is performed to better align the video with the language query.

In the following, we elaborate the main components of the proposed method: (1) feature disentanglement, (2) segment-aware query representation (3) multimodal fusion, (4) cross-guided contrast, (5) overall loss function, and (6) inference.

### A. Feature Disentanglement

The feature disentanglement network is shown in Figure 3. Let $\boldsymbol{X}^i \in \mathbb{R}^{d_V \times T}$ $(i = 1, 2, \cdots, N)$ represent the video segment features extracted from the pre-trained video model, where $d_V$ is the dimension of video segment features, $T$ is the number of segments per video, and $N$ is the mini-batch size. We first obtain the background features $\boldsymbol{B}^i \in \mathbb{R}^{d_V \times T}$ $(i = 1, 2, \cdots, N)$ based on adjacent segments by a background encoder **BE**:

$$\boldsymbol{b}_t^i = \mathbf{BE}(\boldsymbol{x}_t^i, \boldsymbol{x}_{t'}^i), \tag{2}$$

where $\boldsymbol{b}_t^i$, $\boldsymbol{x}_t^i$, and $\boldsymbol{x}_{t'}^i$ are the $t$-th segment feature of $\boldsymbol{B}^i$, the $t$-th segment feature of $\boldsymbol{X}^i$, and the adjacent segment feature of $\boldsymbol{x}_t^i$, respectively.

The action features $\boldsymbol{A}^i \in \mathbb{R}^{d_V \times T}$ are then obtained by an action encoder **AE**:

$$\boldsymbol{a}_t^i = \mathbf{AE}(\boldsymbol{x}_t^i, \boldsymbol{b}_t^i), \tag{3}$$

where $\boldsymbol{a}_t^i$ is the $t$-th segment feature of $\boldsymbol{A}^i$.

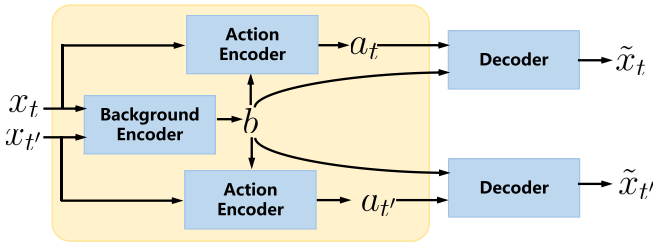



Fig. 3. Feature Disentanglement Network.

A decoder $\mathbf{D}$ is employed to reconstruct the video segment features:

$$\tilde{\boldsymbol{x}}_t^i = \mathbf{D}(\boldsymbol{a}_t^i, \boldsymbol{b}_t^i),$$
$$\tilde{\boldsymbol{x}}_{t'}^i = \mathbf{D}(\mathbf{AE}(\boldsymbol{x}_{t'}^i, \boldsymbol{b}_t^i), \boldsymbol{b}_t^i), \tag{4}$$

where $\tilde{\boldsymbol{x}}_t^i$ and $\tilde{\boldsymbol{x}}_{t'}^i$ are the reconstructed features of $\boldsymbol{x}_t^i$ and its adjacence $\boldsymbol{x}_{t'}^i$.

The reconstructed features are encouraged to be equal to the original segment features and the feature disentanglement loss $\mathcal{L}_{fd}$ can be formulated as follows:

$$\mathcal{L}_{fd} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(||\boldsymbol{x}_t^i - \tilde{\boldsymbol{x}}_t^i||_1 + ||\boldsymbol{x}_{t'}^i - \tilde{\boldsymbol{x}}_{t'}^i||_1). \tag{5}$$

It is notable that the inputs of our disentanglement network are adjacent video segments and the learned background factor is simultaneously employed to reconstruct the video segment and its adjacence. The disentanglement is guaranteed by the inductive bias introduced by adjacent segments, *i.e.*, background factor is enforced to be shared by adjacent segments and action factor is complementary to background one to reconstruct the original video segment.

### B. Segment-Aware Query Representation

For word-embedding features from the pre-trained language model, we first employ a bi-directional LSTM to obtain the word-level query feature $\boldsymbol{W}^i \in \mathbb{R}^{d_Q \times L}$ $(i = 1, 2, \cdots, N)$, where $d_Q$ is the feature dimension and $L$ is the number of words in a sentence. Specifically, the $l$-th query feature in a sentence is obtained by concatenating the hidden states in the forward and backward LSTMs, *i.e*, $\boldsymbol{w}_l^i = [\overrightarrow{h_l}; \overleftarrow{h_l}] \in \mathbb{R}^{d_Q}$.

Following [32], [54], [55], we employ an attention network to obtain the segment-aware query representation $\boldsymbol{Q}^i \in \mathbb{R}^{d_Q \times T}$. To be specific, the query feature at $t$-th position is obtained by attaching different weights for word-level features based on the video segment feature:

$$\boldsymbol{q}_t^i = \sum_{l=1}^{L} \lambda_{tl} \boldsymbol{w}_l^i, \tag{6}$$

where $\lambda_{tl}$ is the normalized attention weight computed by the $t$-th disentangled segment feature, *i.e*, $\lambda_{tl} = \dfrac{\exp(r_{tl})}{\sum_{k=1}^{L}\exp(r_{tk})}$ and $r_{tl} = \boldsymbol{w}_r^T tanh(\mathbf{W_q}\boldsymbol{w}_l^i + \mathbf{W_v}[\boldsymbol{a}_t^i; \boldsymbol{b}_t^i] + \boldsymbol{b}_r)$. $\mathbf{W_q} \in \mathbb{R}^{d \times d_Q}$ and $\mathbf{W_v} \in \mathbb{R}^{d \times 2d_V}$ are the learnable embedding matrices for different modalities, while $\boldsymbol{w}_r^T$ is a trainable vector.

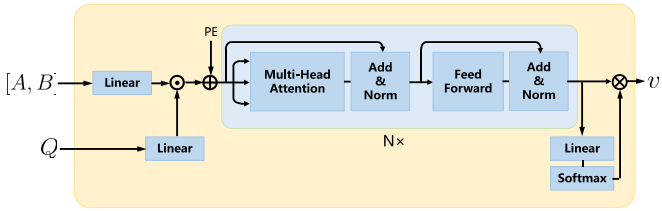



Fig. 4. Multimodal Fusion Network.

### C. Multimodal Fusion

Figure 4 illustrates our multimodal fusion network. We first embed the concatenation of disentangled features, *i.e.*, action features $\boldsymbol{A}^i$ and background features $\boldsymbol{B}^i$, and query feature $\boldsymbol{Q}^i$ into a common space. We then multiply them element by element to obtain the fused video features $\boldsymbol{F}^i$:

$$\boldsymbol{f}_t^i = \mathbf{W_v}[\boldsymbol{a}_t^i, \boldsymbol{b}_t^i] \odot \mathbf{W_q}\boldsymbol{q}_t^i, \tag{7}$$

where $\boldsymbol{f}_t^i$ $(t = 1, 2, \cdots, T)$ is the $t$-th segment feature of $\boldsymbol{F}^i$, $\mathbf{W_v} \in \mathbb{R}^{d \times 2d_V}$, $\mathbf{W_q} \in \mathbb{R}^{d \times d_Q}$ are learnable embedding matrices for different modalities, $\boldsymbol{a}_t^i$ and $\boldsymbol{b}_t^i$ are the $t$-th action and background features of the $i$-th video in the mini-batch, and $\odot$ represents the element-wise multiplication.

Transformer [56] have been shown to successfully capture long-range dependencies for context modeling. Here, in order to capture the contextual and temporal information of video segments, we also adopt a transformer as follows:

$$\boldsymbol{S}^i = \mathbf{TransEncoder}(\boldsymbol{F}^i + \boldsymbol{PE}), \tag{8}$$

where **TransEncoder** and $\boldsymbol{PE}$ are referred to as the transformer encoder and fixed positional encodings, respectively. $\boldsymbol{S}^i \in \mathbb{R}^{d \times T}$ $(i = 1, 2, \cdots, B)$ represents the final video segment features.

In addition, a weighting vector is obtained using another embedding matrix followed by a softmax function:

$$\boldsymbol{m}^i = \mathbf{softmax}(\mathbf{W_s}\boldsymbol{S}^i), \tag{9}$$

where $\mathbf{W_s} \in \mathbb{R}^{1 \times d}$ is another learnable embedding matrix and $\boldsymbol{m}^i \in \mathbb{R}^T$ $(i = 1, 2, \cdots, B)$ is the weighting vector employed to aggregate the video segment features into a video clip representation.

### D. Cross-Guided Contrast

In order to precisely align the video with the query and obtain discriminative features for location regression, we perform cross-guided contrast after the multimodal fusion. Specifically, for the query-guided video contrast, the language query is regarded as an anchor and the video clip corresponding to the query is considered as a positive sample. Video clips corresponding to other queries are negative samples and different weights are attached to negative samples based on the similarities between queries. For video-guided query contrast, the video clip, its corresponding query, and other queries are selected as an anchor, a positive sample and negative samples, respectively. And negative samples are likewise weighted based on the similarities between video clips.

In this section, we first briefly introduce the basic form of contrastive loss. Then, our query-guided video contrast and video-guided query contrast are presented in detail.

*1) Contrastive Loss:* Contrastive learning aims to learn data representations by pulling an anchor closer towards a positive example and pushing the anchor away from many negative examples. Moreover, it can be viewed as a generalization of triplet loss where many negative samples instead of one are employed per anchor. Denote $\boldsymbol{a}$ as an anchor, $\boldsymbol{p}$ as a positive example and $\boldsymbol{n}_i$ $(i = 1, 2, \cdots, K-1)$ as negative examples. The contrastive loss can be formulated as:

$$\mathcal{L} = -\log \frac{e^{\mathbf{s}(\boldsymbol{a},\boldsymbol{p})}}{e^{s(\boldsymbol{a},\boldsymbol{p})} + \sum_{i=1}^{K-1} e^{s(\boldsymbol{a},\boldsymbol{n}_i)}}, \tag{10}$$

where $\mathbf{s}(\cdot,\cdot)$ is a similarity function and $K$ is the number of contrastive samples per anchor.

For unsupervised contrastive learning [51], positive examples are generated with data augmentation, and negative examples are randomly chosen from other examples in the mini-batch. While, for supervised VTG task, where labels are available during training process, we need to generalize the self-supervised contrastive learning to supervised settings, in order to fully leverage the label information.

*2) Query-Guided Video Contrast:* Let $\tilde{\boldsymbol{m}}^i \in \mathbb{R}^T$ $(i = 1, 2, \cdots, N)$ be segment masks obtained from the ground-truth. $\tilde{\boldsymbol{m}}_t^i$ $(t = 1, 2, \cdots, T)$ is set to 1 if the $t$-th segment is located within the ground-truth time interval and 0 otherwise. The global representation of query $\boldsymbol{q}^i$ and video clip $\boldsymbol{v}^i$ are obtained from the segment-aware query features $\boldsymbol{Q}^i$ and video segment features $\boldsymbol{S}^i$ pooled by the segment masks:

$$\begin{aligned} \boldsymbol{q}^i &= \sum_{t=1}^{T} \tilde{\boldsymbol{m}}_t^i \boldsymbol{Q}_t^i, \\ \boldsymbol{v}^i &= \sum_{t=1}^{T} \tilde{\boldsymbol{m}}_t^i \boldsymbol{S}_t^i. \end{aligned} \tag{11}$$

For video contrastive learning, we choose video clips corresponding to other queries as negative samples. As shown in Eq.(10), vanilla contrastive loss treats all the negative samples equally; this influences the representation learning as the negative samples are likely to be semantically similar to the anchor. We therefore attach different weights to negative video samples based on the similarities between queries and the query-guided video contrast loss can be defined as:

$$\mathcal{L}_{cc\_video} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathbf{s}(\boldsymbol{q}^i,\boldsymbol{v}^i)}}{e^{\mathbf{s}(\boldsymbol{q}^i,\boldsymbol{v}^i)} + \sum_{k\neq i}^{N} \mathbf{w}(\boldsymbol{q}^i,\boldsymbol{q}^k) e^{\mathbf{s}(\boldsymbol{q}^i,\boldsymbol{v}^k)}}, \tag{12}$$

where $\mathbf{w}(\cdot,\cdot)$ is the weighting function and $N$ is the size of mini-batch. $\mathbf{w}(\boldsymbol{x}_1,\boldsymbol{x}_2) = 1 - \dfrac{\boldsymbol{x}_1\boldsymbol{x}_2}{||\boldsymbol{x}_1||_2||\boldsymbol{x}_2||_2}$ is used in our paper. In the extreme scenarios where two queries are semantically equal, the corresponding video clip is not treated as a negative sample.

*3) Video-Guided Query Contrast:* We further present video-guided query contrast to obtain discriminative language representation and the video-guided query contrast loss is defined likewise:

$$\mathcal{L}_{cc\_query} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\mathbf{s}(\boldsymbol{v}^i,\boldsymbol{q}^i)}}{e^{\mathbf{s}(\boldsymbol{v}^i,\boldsymbol{q}^i)} + \sum_{k\neq i}^{N} \mathbf{w}(\boldsymbol{v}^i,\boldsymbol{v}^k) e^{\mathbf{s}(\boldsymbol{v}^i,\boldsymbol{q}^k)}}. \tag{13}$$

The cross-guided contrast loss consists of two parts:

$$\mathcal{L}_{cc} = \mathcal{L}_{cc\_video} + \mathcal{L}_{cc\_query}. \tag{14}$$

### *E. Overall Loss Function*

In addition to the feature disentanglement loss and cross-guided contrast loss, a temporal weighting loss $\mathcal{L}_{tw}$ and a grounding loss $\mathcal{L}_{grd}$ are also adopted in our method to align the time interval with the ground-truth. Specifically, the temporal weighting loss makes the video segments within ground-truth time interval with higher weighting values:

$$\mathcal{L}_{tw} = -\frac{\sum_{t=1}^{T} \tilde{\boldsymbol{m}}_t^i \log(\boldsymbol{m}_t^i)}{\sum_{t=1}^{T} \tilde{\boldsymbol{m}}_t^i}, \tag{15}$$

where $\boldsymbol{m}_j^i$ and $\tilde{\boldsymbol{m}}_j^i$ are the weighting vector and its ground-truth counterpart as mentioned before. The grounding loss is defined as the sum of $\mathbf{smooth}\ell_\mathbf{1}$ distances between the normalized ground-truth time interval $((\tilde{t}^s, \tilde{t}^e) \in [0,1])$ and our prediction $(t^s, t^e)$:

$$\mathcal{L}_{grd} = \mathbf{smooth}\ell_\mathbf{1}(\tilde{t}^s - t^s) + \mathbf{smooth}\ell_\mathbf{1}(\tilde{t}^e - t^e), \tag{16}$$

where $\mathbf{smooth}\ell_\mathbf{1}(x) = 0.5x^2$ if $|x| < 1$,and $\mathbf{smooth}\ell_\mathbf{1}(x) = |x| - 0.5$ otherwise.

In summary, four loss functions are employed for parameter optimization, namely: temporal weighting loss $\mathcal{L}_{tw}$, grounding loss $\mathcal{L}_{grd}$, feature disentanglement loss $\mathcal{L}_{fd}$, and cross-guided contrast loss $\mathcal{L}_{cc}$. The overall objective function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{tw} + \alpha\mathcal{L}_{grd} + \beta\mathcal{L}_{fd} + \gamma\mathcal{L}_{cc}, \tag{17}$$

where $\alpha$, $\beta$, and $\gamma$ are the weighting coefficients for the grounding loss, the feature disentanglement loss, and the cross-guided contrast loss, respectively.

### *F. Inference*

During the inference period, the video segment features $\boldsymbol{X}$ extracted from the pre-trained model are first disentangled into action factors and background factors by our feature disentanglement network. Subsequently, the disentangled video action features $\boldsymbol{A}$, video background features $\boldsymbol{B}$, and language query feature $\boldsymbol{Q}$ from the attention module are fed into the multimodal fusion network to generate the fused segment features $\boldsymbol{S}$ and the weighting vector $\boldsymbol{m}$, as shown in Equation (7), (8), and (9). The time interval is next regressed based on the weighted video features $\boldsymbol{v} = \sum_{t=1}^{T} \boldsymbol{m}_t \boldsymbol{S}_t$. It should be noted that the learned weighting vectors $\boldsymbol{m}$ from the multimodal fusion module are used for inference, while the weighting vectors $\tilde{\boldsymbol{m}}$ obtained from the ground-truth are directly employed to generate contrastive examples during training.

## IV. Experiments

In this section, the datasets and evaluation metrics are first introduced in detail. We then present the implementation details and the experimental results of our method compared against other several state-of-the-art methods. Finally, the ablation studies and qualitative results further prove the effectiveness of our proposed approach.

### A. Datasets

We evaluate the proposed method on three public datasets: TACoS [57], Charades-STA [22], and ActivityNet Captions [58].

*1) TACoS:* The TACoS dataset is built on the MPII-Compositive dataset [59] and consists of 10,146, 4,589, and 4,083 instances for training, validation, and testing, respectively.

*2) Charades-STA:* The Charades-STA dataset is built on the Charades dataset for video temporal grounding. It contains 12,408 and 3,720 clip-language pairs for training and testing, respectively. The average length of the language query, average video duration and average number of activities per video are 8.6 words, 29.8 seconds, and 2.3 respectively.

*3) ActivityNet Captions:* The ActivityNet Captions dataset contains 10,024, 4,926 and 5,044 videos for training, validation, and testing, respectively. The average length of language query and average number of activities per video are 13.48 words and 3.65, respectively. We use the validation set for evaluation since the testing set is not publicly available.

### B. Evaluation Metrics

Following previous works, we report two metrics to measure the performance of video temporal grounding: Rank@1 IoU=n and mIoU.

*1) Rank@1 IoU=n:* Rank@1 IoU=n is referred to as the percentage of test samples whose Intersection over Union (IoU) with ground-truth (GT) is higher than $n$. $n = \{0.3, 0.5, 0.7\}$ are reported in our experiments.

**mIoU.** mIoU is the mean IoU over all test samples.

### C. Implementation Details

*1) Video Segment and Language Query Features:* We use pre-trained C3D [60], I3D [61], and C3D models to extract video features for the TACoS, Charades-STA and ActivityNet Captions datasets, respectively. Each video is uniformly sampled into $T = 128$ segments and embedded into a $d_V \times T$ dimensional representation space by an embedding matrix. Glove [62] is employed to extract the word features of the language query. The dimensions of the video segment $d_V$, the language query $d_Q$, and the common space $d$ are set to 1024, 512, and 512, respectively.

*2) Optimization:* Our approach is implemented with PyTorch [63] and optimized by ADAM [64] optimizer with a learning rate of 0.0004. The loss weights $\alpha$, $\beta$, and $\gamma$ are determined via grid-search and set to 1, 10, and 0.5, respectively. In addition, we set the batch size $N$ to 64.

*3) Other Details:* We use MLP as the background encoder **BE**, the action encoder **AE**, and the decoder **D** in the feature disentanglement network. The number of stacked layers and heads in transformer encoder are set to 1 and 8, respectively. Cosine similarity is employed as the similarity function $\mathbf{s}(\cdot, \cdot)$ in cross-guided contrast.

### D. Comparison With State-of-the-Art Methods

*1) Compared Methods:* We compare the proposed method with several state-of-the-art methods. Among these methods, CTRL [22], SAP [23], MLVI [65], MAN [66], SCDM [67], and 2D-TAN [29] are proposal-based methods. SM-RL [33], RWM [28], TripNet [27], and TSP-PRL [34] are reinforcement learning-based methods. ABLR [24], GDP [30], VSLNet [68], LGI [25], DRN [69], PMI-LOC [31], IVG-DCL [26], and ACRM [32] are proposal-free methods. The details of different methods are summarized below.

- CTRL is a pioneering VTG work. It leverages sliding window to obtain candidate clips of various lengths and fuse the candidate representations with the sentence representation by three operators (*i.e.*,add, multiply, and fully-connected layer) to predict the alignment score.
- SAP proposes a Semantic Activity Proposal framework that integrates the semantic information of sentence queries into the proposal generation process to get discriminative activity proposals. Visual and semantic information are jointly utilized for proposal ranking and refinement.
- MLVI introduces a multilevel feature integration model to fuse language and vision earlier and more tightly.
- MAN presents a Moment Alignment Network where language query is integrated as dynamic filters. In addition, an iterative graph adjustment network is devised to model moment-wise temporal relations.
- SCDM proposes a semantic conditioned dynamic modulation mechanism for VTG, where language query is employed to modulate the temporal convolution operations for better correlating and composing the sentence related video contents over time.
- 2D-TAN proposes to model the temporal relations between video candidates by a two-dimensional map, where one dimension indicates the starting time of a moment and the other indicates the end time. The 2D temporal map can cover diverse video moments with different lengths, while representing their adjacent relations.
- SM-RL proposes a recurrent neural network based reinforcement learning model which selectively observes a sequence of frames and associates the given sentence with video content in a matching-based manner.
- RWM views VTG as controlling an agent to read the description, to watch the video as well as the current localization, and then to move the temporal grounding boundaries iteratively to find the best matching clip.
- TripNet introduces an end-to-end reinforcement learning framework that uses a gated-attention mechanism over cross-modal features.
- TSP-PRL presents a Tree-Structured Policy based Progressive Reinforcement Learning framework to sequentially regulate the temporal boundary by an iterative refinement process.

TABLE I
PERFORMANCE COMPARISIONS ON THE TACOS, CHARADES-STA, AND ACTIVITYNET CAPTIONS DATASETS

| Method | TACoS | | | | Charades-STA | | | | ActivityNet Captions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU |
| CTRL | 0.1832 | 0.1330 | - | - | - | 0.2363 | 0.0889 | - | 0.2870 | 0.1400 | - | 0.2054 |
| SAP | - | 0.1824 | - | - | - | 0.2742 | 0.1336 | - | - | - | - | - |
| MLVI | - | - | - | - | 0.5470 | 0.3560 | 0.1580 | - | 0.4530 | 0.2770 | 0.1360 | - |
| MAN | - | - | - | - | - | 0.4653 | 0.2272 | - | - | - | - | - |
| SCDM | 0.2611 | 0.2117 | - | - | - | 0.5444 | 0.3343 | - | 0.5480 | 0.3675 | 0.1986 | - |
| 2D-TAN | 0.3729 | 0.2532 | - | - | - | - | - | - | - | - | - | - |
| SM-RL | 0.2025 | 0.1595 | - | - | - | 0.2436 | 0.1117 | - | - | - | - | - |
| TripNet | - | - | - | - | 0.5133 | 0.3661 | 0.1450 | - | 0.4842 | 0.3219 | 0.1393 | - |
| RWM | - | - | - | - | - | 0.3670 | - | - | - | 0.3690 | - | - |
| TSP-PRL | - | - | - | - | - | 0.4545 | 22475 | - | 0.5602 | 0.3882 | - | - |
| ABLR | 0.1950 | 0.094 | - | - | - | - | - | - | 0.5567 | 0.3679 | - | 0.3699 |
| DEBUG | 0.2345 | 0.1172 | - | 0.1603 | - | - | - | - | - | - | - | - |
| GDP | - | - | - | - | 0.5454 | 0.3947 | 0.1849 | - | 0.5617 | 0.3927 | - | 0.3980 |
| VSLNet | 0.2961 | 0.2427 | 0.2003 | 0.2411 | 0.7046 | 0.5419 | 0.3522 | 0.5002 | 0.6316 | 0.4322 | 0.2616 | 0.4319 |
| LGI | - | - | - | - | 0.7296 | 0.5946 | 0.3548 | 0.5138 | 0.5852 | 0.4151 | 0.2307 | 0.4113 |
| DRN | - | 0.2317 | - | - | - | 0.5309 | 0.3175 | - | - | 0.4545 | 0.2436 | - |
| PMI-LOC | - | - | - | - | 0.5548 | 0.3973 | 0.1927 | - | 0.5969 | 0.3828 | 0.1783 | - |
| IVG-DCL | 0.3884 | 0.2907 | 0.1905 | 0.2826 | 0.6763 | 0.5024 | 0.3288 | 0.4802 | 0.6322 | 0.4384 | 0.2710 | 0.4421 |
| ACRM | 0.5119 | 0.3879 | 0.2694 | 0.3742 | 0.7347 | 0.5753 | 0.3833 | 0.5301 | - | - | - | - |
| Ours | **0.5331** | **0.4214** | **0.2932** | **0.3885** | **0.7484** | **0.6110** | **0.3970** | **0.5357** | **0.6458** | **0.4536** | **0.2768** | **0.4545** |

TABLE II
IMPORTANCE OF THE TRANSFORMER BASED MULTIMODAL FUSION ON THE TACOS, CHARADES-STA, AND ACTIVITYNET CAPTIONS DATASETS. ENCODER AND PE ARE REFERRED TO AS TRANSFORMER ENCODER AND POSITIONAL ENCODING, RESPECTIVELY

| Method | TACoS | | | | Charades-STA | | | | ActivityNet Captions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU |
| Base Model | **0.5111** | **0.3874** | **0.2604** | **0.3705** | **0.7296** | **0.5830** | **0.3804** | **0.5229** | **0.6205** | **0.4394** | **0.2552** | **0.4348** |
| w/o Encoder | 0.4586 | 0.3489 | 0.2359 | 0.3364 | 0.7003 | 0.5016 | 0.3291 | 0.4936 | 0.5987 | 0.3979 | 0.2150 | 0.4103 |
| w/o PE | 0.4911 | 0.3697 | 0.2509 | 0.3562 | 0.7129 | 0.5347 | 0.3446 | 0.5072 | 0.6085 | 0.4079 | 0.2236 | 0.4255 |

- ABLR proposes a Attention Based Location Regression model to directly regress the temporal coordinates from the global attention outputs with a multi-modal co-attention mechanism.
- GDP designs a novel bottom-up model: Graph-FPN with Dense Predictions. It first generates a frame feature pyramid to capture multi-level semantics, then utilizes graph convolution to encode the plentiful scene relationships.
- VSLNet proposes a video span localizing network on top of the standard span-based QA framework with query-guided highlighting strategy.
- LGI introduces a sequential query attention module to extract representations of multiple and distinct semantic phrases from a text query. Then Local-Global video-text Interaction algorithm is employed to model the relationship between video segments and semantic phrases in multiple levels.
- DRN proposes a Dense Regression Network for VTG, which provides a new perspective to leverage dense supervision from the sparse annotations.
- PMI-LOC proposes pairwise modality interaction in both the sequence and channel levels to better understand video contents.
- IVG-DCL proposes interventional video grounding to eliminate the spurious correlations between query and video features based on causal inference. In addition, a dual contrastive learning approach is employed to better align the text and video.
- ACRM propose an Attentive Cross-modal Relevance Matching model which introduces an attention mechanism to model the interactions between the video frame and query word features.

*2) Performance Analysis:* As shown in Table I, our method achieves state-of-the-art performances on the TACoS, Charades-STA, and ActivityNet Captions datasets. For example, our proposed method can surpass the previous best approach ACRM by 2.12%, 3.35%, 2.38%, and 1.43% in terms of Rank@1 IoU=0.3, Rank@1 IoU=0.5, Rank@1 IoU=0.7, and mIoU respectively on the TACoS dataset. The results indicates the effectiveness of our feature disentanglement, transformer based multimodal fusion, and cross-guided contrast.

### *E. Ablation Studies*

*1) Importance of the Transformer Based Multimodal Fusion:* To evaluate the importance of the components of our multimodal fusion network, we vary our base model in two different ways. Specifically, w/o Encoder and w/o PE are referred to as our multimodal fusion module with transformer encoder or positional encodings removed, respectively.

TABLE III

GAINS FROM FEATURE DISENTANGLEMENT ON THE TACOS, CHARADES-STA, AND ACTIVITYNET CAPTIONS DATASETS. $\boldsymbol{x}$, $\boldsymbol{a}$, AND $\boldsymbol{b}$ REPRESENT THE ENTANGLED SEGMENT FEATURE, DISENTANGLED ACTION FEATURE, AND DISENTANGLED BACKGROUND FEATURE, RESPECTIVELY

| Method | TACoS | | | | Charades-STA | | | | ActivityNet Captions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU |
| $\boldsymbol{x}$ | 0.5111 | 0.3874 | 0.2604 | 0.3705 | 0.7296 | 0.5830 | 0.3804 | 0.5229 | 0.6205 | 0.4394 | 0.2552 | 0.4348 |
| $\boldsymbol{a}$ | 0.5216 | 0.3982 | 0.2797 | 0.3794 | 0.7321 | 0.5976 | 0.3917 | 0.5283 | 0.6294 | 0.4452 | 0.2624 | 0.4435 |
| $\boldsymbol{b}$ | 0.5129 | 0.3907 | 0.2759 | 0.3761 | 0.7306 | 0.5907 | 0.3864 | 0.5255 | 0.6270 | 0.4410 | 0.2596 | 0.4397 |
| $[\boldsymbol{a}, \boldsymbol{b}]$ | **0.5221** | **0.4108** | **0.2802** | **0.3806** | **0.7363** | **0.6003** | **0.3954** | **0.5305** | **0.6351** | **0.4481** | **0.2667** | **0.4462** |

TABLE IV

CONTRIBUTION OF THE DIFFERENT CONTRAST LOSSES ON THE TACOS, CHARADES-STA, AND ACTIVITYNET CAPTIONS DATASETS. $+\mathcal{L}_{cc\_video}$, $+\mathcal{L}_{cc\_query}$, $+\mathcal{L}_{cc}$, AND $+\mathcal{L}_{con}$ ARE REFERRED AS QUERY-GUIDED VIDEO CONTRAST, VIDEO-GUIDED QUERY CONTRAST, CROSS-GUIDED CONTRAST, AND $\mathcal{L}_{cc}$ WITHOUT CROSS-GUIDED WEIGHTING, RESPECTIVELY

| Method | TACoS | | | | Charades-STA | | | | ActivityNet Captions | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU | Rank@1 IoU=0.3 | Rank@1 IoU=0.5 | Rank@1 IoU=0.7 | mIoU |
| Base Model | 0.5221 | 0.4108 | 0.2802 | 0.3806 | 0.7363 | 0.6003 | 0.3954 | 0.5305 | 0.6351 | 0.4481 | 0.2667 | 0.4462 |
| $+\mathcal{L}_{cc\_video}$ | 0.5269 | 0.4119 | 0.2809 | 0.3824 | 0.7435 | 0.6019 | 0.3955 | 0.5321 | 0.6403 | 0.4526 | 0.2718 | 0.4512 |
| $+\mathcal{L}_{cc\_query}$ | 0.5256 | 0.4121 | 0.2834 | 0.3826 | 0.7422 | 0.6054 | 0.3895 | 0.5311 | 0.6394 | 0.4495 | 0.2688 | 0.4507 |
| $+\mathcal{L}_{cc}$ | **0.5331** | **0.4214** | **0.2932** | **0.3885** | **0.7484** | **0.6110** | **0.3970** | **0.5357** | **0.6458** | **0.4536** | **0.2768** | **0.4545** |
| $+\mathcal{L}_{con}$ | 0.5299 | 0.4141 | 0.2779 | 0.3848 | 0.7384 | 0.6032 | 0.3915 | 0.5312 | 0.6408 | 0.4537 | 0.2713 | 0.4496 |

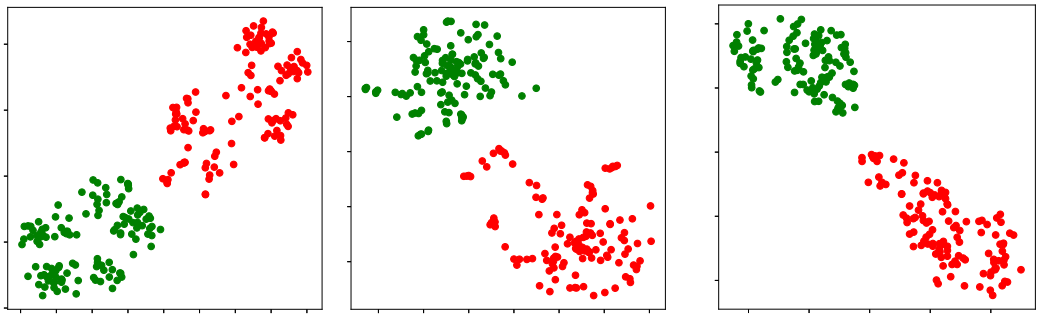


Fig. 5. Visualization of the disentangled action and background features using t-SNE on the TACoS dataset. Red and green points represent the disentangled action and background features, respectively.

The changes in performance on the TACoS, Charades-STA, and ActivityNet Captions datasets are presented in Table II. We can observe that the grounding performance degrades compared to the base model when the encoder or positional encodings removed. This indicates that of our transformer-based multimodal fusion method can effectively capture contextual and temporal information.

*2) Gains From the Feature Disentanglement:* In order to investigate the gains from our feature disentanglement network, we conduct experiments using different video segment features on the TACoS, Charades-STA, and ActivityNet Captions datasets. Four features are employed for comparison, *i.e.*, entangled segment feature $\boldsymbol{x}$, disentangled action feature $\boldsymbol{a}$, disentangled background feature $\boldsymbol{b}$, and the concatenation of disentangled features $[\boldsymbol{a}, \boldsymbol{b}]$.

As shown in Table III, replacing the entangled segment feature with the disentangled action feature can improve the grounding performance by a large margin. It is notable that background factors can also yield improvements. We attribute this result to that action factors and background factors are complementary and correspond to different parts of a language query. The best performance is achieved when action factors and background factors are used together.

*3) Contribution of Cross-Guided Contrast:* We next perform an ablation study to validate the contribution of our contrast losses on the TACoS, Charades-STA, and ActivityNet Captions datasets. Five variants of our model are trained in this experiment: Base Model, $+\mathcal{L}_{cc\_video}$, $+\mathcal{L}_{cc\_query}$, $+\mathcal{L}_{cc}$, and $+\mathcal{L}_{con}$. Specifically, Base Model denotes that only the temporal weighting loss $\mathcal{L}_{tw}$, grounding loss $\mathcal{L}_{grd}$, and feature disentanglement loss $\mathcal{L}_{fd}$ are used for training. $+\mathcal{L}_{cc\_video}$ and $+\mathcal{L}_{cc\_query}$ represent the Base Model with addition of query-guided video contrast loss and video-guided query loss, respectively. $+\mathcal{L}_{cc}$ is our full model where temporal weighting loss $\mathcal{L}_{tw}$, grounding loss $\mathcal{L}_{grd}$, feature disentanglement loss $\mathcal{L}_{fd}$, and cross-guided contrast loss $\mathcal{L}_{cc}$ are all employed for training. In addition, $+\mathcal{L}_{con}$ is referred to as the addition of contrast loss without cross-guided weighting, *i.e.*, negative samples are treated equally.

Table IV presents the experimental results. As can be seen from the table, whether used independently or together, our proposed query-guided video contrast and video-guided query

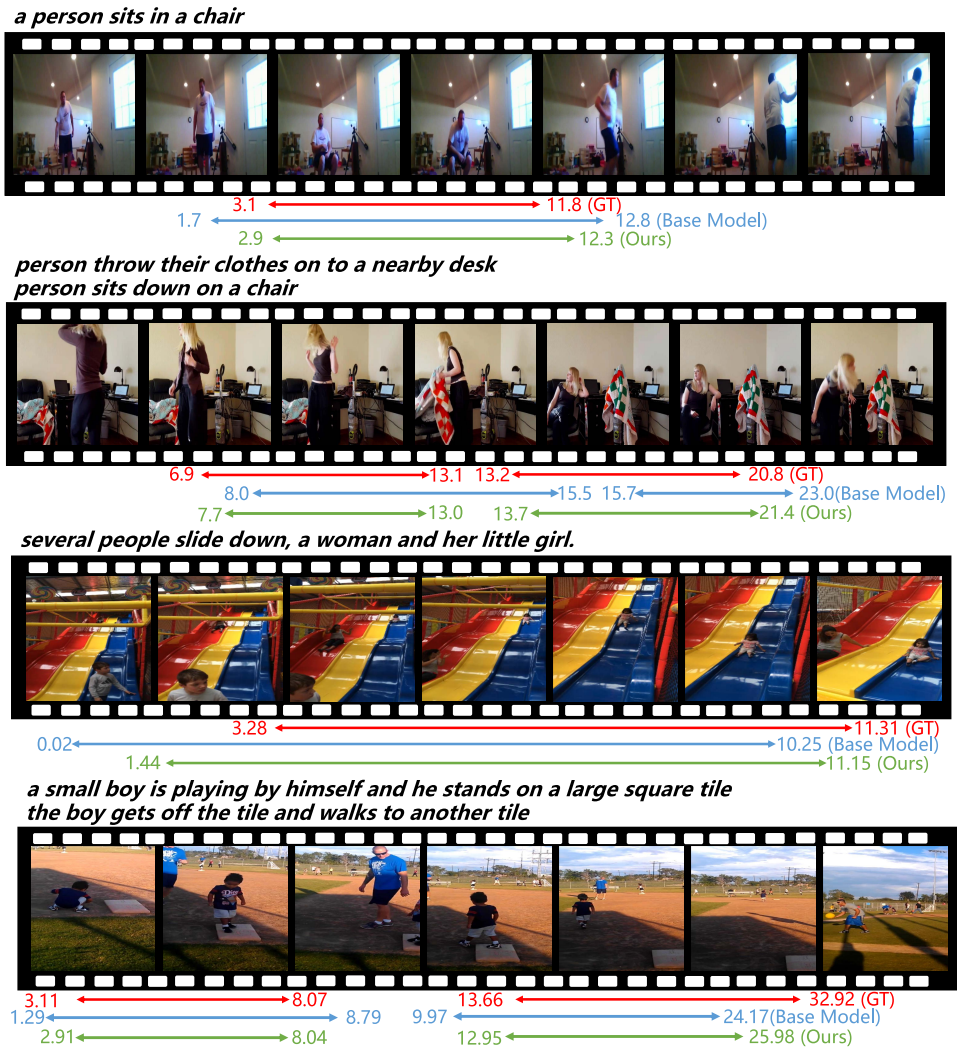



Fig. 6. The qualitative results of different models on Charades-STA and ActivityNet Captions datasets. GT is the ground-truth time interval. Ours represents the proposed HiSA model and Base Model is refered to as HiSA without feature disentanglement and cross-guided contrast.

contrast bring about considerable improvements. In addition, the last two rows of the table show that assigning different weights for negative samples improves the grounding performance. The results demonstrate the effectiveness of our proposed cross-guided contrast, indicating that we have achieved a better alignment of video and language query.

*4) Performance of the Proposed Module on Other Video Understanding Tasks:* In order to investigate whether the proposed method can boost other video understanding tasks, we conduct experiments for a widely studied task called weakly-supervised temporal action localization (WSTAL). Specifically, given an untrimmed video, WSTAL aims to simultaneously locate the time interval and recognize the categories of pre-defined actions with only video-level action labels available during training. We add the proposed feature disentanglement module to ASL [70], which is one of the state-of-the-art WSTAL methods.

Following the standard protocol on temporal action localization, we evaluate the methods with mean Average Precision (mAP) under different Intersection-over-Union (IoU) thresholds. Table V shows the performance on the THUMOS14 [71] dataset. ASL+FD can significantly outperform the baseline method with the addition of feature disentanglement, which verifies the effectiveness of the proposed method.

TABLE V

PERFORMANCE COMPARISON OF THE WEAKLY-SUPERVISED TEMPORAL ACTION LOCALIZATION TASK ON THE THUMOS14 DATASET

| Method | mAP(%)@IoU | | | | | AVG |
|---|---|---|---|---|---|---|
| | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 | (0.1:0.9) |
| ASL | 67.0 | 51.8 | 31.1 | 11.4 | **0.7** | 32.2 |
| ASL+FD | **69.5** | **55.0** | **33.8** | **12.7** | **0.7** | **34.2** |

## F. Qualitative Results

*1) Visualization of the Disentangled Features:* In order to verify that different aspects of video segments can be captured by the feature disentanglement network, we randomly sample

videos and visualize the learned action and background factors of video segments using t-SNE [72]. As shown in Figure 5, the action and background factors are distinctly distributed in the feature space, which indicates the effectiveness of the proposed feature disentanglement approach.

*2) Qualitative Examples of Different Models:* Figure 6 demonstrates the qualitative results of different models. Our proposed model HiSA achieves more accurate video temporal grounding than the Base Model without feature-disentanglement and cross-guided contrast, which demonstrates that our method can produce better alignment between the video and language query.

## V. Conclusion and Future Work

In this paper, we propose Hierarchically Semantic Associating (HiSA) for video temporal grounding, which can better align the video with language query by jointly considering the intra-video entanglement and inter-video connection. Based on adjacent segment features, we disentangle the video segment into action factors and background factors, providing a novel solution for better multimodal interaction. In addition, cross-guided contrast is framed to establish inter-video connection. State-of-the-art performances are achieved on the TACoS, Charades-STA, and ActivityNet Captions datasets.

Existing fully-supervised VTG methods require accurate annotations of temporal boundary, which is time-consuming and expensive to obtain. In the future, we will explore to solve the VTG problem under the weakly-supervised setting, where only video and language pairs are available during training.

## References


[1] N. Vaswani and R. Chellappa, "Principal components null space analysis for image and video classification," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1816–1830, Jul. 2006.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. CVPR*, Jun. 2014, pp. 1725–1732.

[3] F. Yu, X. Wu, J. Chen, and L. Duan, "Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5308–5321, Nov. 2019.

[4] C. Gan, C. Sun, L. Duan, and B. Gong, "Webly-supervised video recognition by mutually voting for relevant web images and web video frames," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 849–866.

[5] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, "You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images," in *Proc. CVPR*, Jun. 2016, pp. 923–932.

[6] M. Kristan *et al.*, "The visual object tracking VOT2015 challenge results," in *Proc. ICCVW*, 2015, pp. 1–23.

[7] R. Zeng, C. Gan, P. Chen, W. Huang, Q. Wu, and M. Tan, "Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5797–5808, Dec. 2019.

[8] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han, "Revisiting anchor mechanisms for temporal action localization," *IEEE Trans. Image Process.*, vol. 29, pp. 8535–8548, 2020.

[9] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. CVPR*, Jun. 2016, pp. 1049–1058.

[10] D. Guo, W. Li, and X. Fang, "Fully convolutional network for multiscale temporal action proposals," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3428–3438, Dec. 2018.

[11] Z. Zhou, F. Shi, and W. Wu, "Learning spatial and temporal extents of human actions for action detection," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 512–525, Apr. 2015.

[12] R. Zeng *et al.*, "Graph convolutional networks for temporal action localization," in *Proc. ICCV*, Oct. 2019, pp. 7094–7103.

[13] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Proc. CVPR*, 2015, pp. 2568–2577.

[14] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[15] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proc. CVPR*, Jun. 2018, pp. 7622–7631.

[16] J. Lei *et al.*, "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in *Proc. CVPR*, Jun. 2021, pp. 7331–7341.

[17] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," 2019, *arXiv:1904.11574*.

[18] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "STAR: A benchmark for situated reasoning in real-world videos," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–13.

[19] Z. Chen, J. Mao, J. Wu, K.-Y. K. Wong, J. B. Tenenbaum, and C. Gan, "Grounding physical concepts of objects and events through dynamic visual reasoning," in *Proc. ICLR*, 2021, pp. 1–20.

[20] M. Ding, Z. Chen, T. Du, P. Luo, J. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," in *Proc. NeurIPS*, vol. 34, 2021, pp. 887–899.

[21] K. Yi *et al.*, "CLEVRER: Collision events for video representation and reasoning," in *Proc. ICLR*, 2019, pp. 1–19.

[22] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. ICCV*, Oct. 2017, pp. 5267–5275.

[23] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proc. AAAI*, vol. 33, 2019, pp. 8199–8206.

[24] 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*, vol. 33, 2019, pp. 9159–9166.

[25] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. CVPR*, Jun. 2020, pp. 10810–10819.

[26] G. Nan *et al.*, "Interventional video grounding with dual contrastive learning," in *Proc. CVPR*, Jun. 2021, pp. 2765–2775.

[27] M. Hahn, A. Kadav, J. M. Rehg, and H. P. Graf, "Tripping through time: Efficient localization of activities in videos," 2019, *arXiv:1904.09936*.

[28] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," in *Proc. AAAI*, vol. 33, 2019, pp. 8393–8400.

[29] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D temporal adjacent networks for moment localization with natural language," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 12870–12877.

[30] L. Chen *et al.*, "Rethinking the bottom-up framework for query-based video localization," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 10551–10558.

[31] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, "Learning modality interaction for temporal sentence localization and event captioning in videos," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 333–351.

[32] 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.

[33] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proc. CVPR*, Jun. 2019, pp. 334–343.

[34] J. Wu, G. Li, S. Liu, and L. Lin, "Tree-structured policy based progressive reinforcement learning for temporally language grounding in video," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 12386–12393.

[35] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[36] F. Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. ICML*, 2019, pp. 4114–4124.

[37] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 172–189.

[38] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. ECCV*, 2018, pp. 35–51.

[39] S. Jeong, Y. Kim, E. Lee, and K. Sohn, "Memory-guided unsupervised image-to-image translation," 2021, *arXiv:2104.05170*.

[40] X. Li *et al.*, "Image-to-image translation via hierarchical style disentanglement," 2021, *arXiv:2103.01456*.

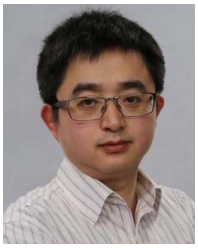
[41] S. Lee, S. Cho, and S. Im, "DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation," 2021, *arXiv:2103.13447*.
[42] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. CVPR*, Jul. 2017, pp. 1415–1424.
[43] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," 2016, *arXiv:1606.03657*.
[44] H. Kim and A. Mnih, "Disentangling by factorising," 2018, *arXiv:1802.05983*.
[45] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "CausalVAE: Structured causal disentanglement in variational autoencoder," 2020, *arXiv:2004.08697*.
[46] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[47] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3VAE: Self-supervised sequential VAE for representation disentanglement and data generation," in *Proc. CVPR*, Jun. 2020, pp. 6538–6547.
[48] S. Ding *et al.*, "Motion-aware contrastive video representation learning via foreground-background merging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 9716–9726.
[49] P. Chen *et al.*, "RSPNet: Relative speed perception for unsupervised video representation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1045–1053.
[50] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2019, *arXiv:1906.05849*.
[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, Jun. 2020, pp. 9729–9738.
[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
[53] P. Khosla *et al.*, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.
[54] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *Proc. ICCV*, Oct. 2017, pp. 1811–1820.
[55] Z. Chen, L. Ma, W. Luo, and K.-Y.-K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," in *Proc. ACL*, 2019, pp. 1884–1894.
[56] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.
[57] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.
[58] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. ICCV*, Oct. 2017, pp. 706–715.
[59] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proc. ECCV*. Berlin, Germany: Springer, 2012, pp. 144–157.
[60] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, Dec. 2015, pp. 4489–4497.
[61] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. CVPR*, Jul. 2017, pp. 6299–6308.
[62] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.
[63] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8026–8037.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[65] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proc. AAAI*, vol. 33, 2019, pp. 9062–9069.
[66] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proc. CVPR*, Jun. 2019, pp. 1247–1257.
[67] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," 2019, *arXiv:1910.14303*.
[68] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," 2020, *arXiv:2004.13931*.
[69] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan, "Dense regression network for video grounding," in *Proc. CVPR*, Jun. 2020, pp. 10287–10296.
[70] J. Ma, S. K. Gorti, M. Volkovs, and G. Yu, "Weakly supervised action selection learning in video," in *Proc. CVPR*, 2021, pp. 7587–7596.
[71] H. Idrees *et al.*, "The THUMOS challenge on action recognition for videos 'in the wild,'" *Comput. Vis. Image Understand.*, vol. 155, pp. 1–23, Feb. 2017.
[72] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal Of Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

**Zhe Xu** received the B.E. degree from Xidian University, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests include computer vision and machine learning.

**Da Chen** received the B.S. degree in optic science from Xidian University, Xi'an, China, the M.S. degree in computer science from Halmstad University, Halmstad, Sweden, and the Ph.D. degree in computer science from the University of Bath, Bath, U.K. He is currently a Researcher with Alibaba Group, Beijing China. His research interests include machine learning, computer vision, and multimodal analysis.

**Kun Wei** received the B.E. and Ph.D. degrees in electronic and information engineering from Xidian University, China, in 2017 and 2022, respectively. He is currently a Lecturer with the School of Electronic Engineering, Xidian University. His research interests include computer vision and machine learning.

**Cheng Deng** (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He is the author or the coauthor of more than 100 scientific articles at top venues, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, ICCV, AAAI, IJCAI, and KDD. His research interests include computer vision, pattern recognition, and information hiding.



**Hui Xue** received the Ph.D. degree from Zhejiang University. He is currently a Senior Staff Engineer with Alibaba Group, and the Head of Alibaba Security Perception and Cognitive Intelligence & Alibaba AI Safety Management. He leads the Alibaba Turing Security Laboratory, Alibaba's Security Department.