WINNER: Weakly-supervised hIerarchical decompositioN and aligNment for spatio-tEmporal video gRounding

Mengze Li [1 *] Han Wang [1*] Wenqiao Zhang [2] Jiaxu Miao [1] Zhou Zhao [1,3,4†]
Shengyu Zhang [1†] Wei Ji [2†] Fei Wu [3,4,1]
[1] Zhejiang University, [2] National university of Singapore,
[3] Shanghai Institute for Advanced Study of Zhejiang University, [4] Shanghai AI Laboratory
{mengzeli, zhouzhao, sy_zhang}@zju.edu.cn weiji0523@gmail.com

Abstract

Spatio-temporal video grounding aims to localize the aligned visual tube corresponding to a language query. Existing techniques achieve such alignment by exploiting dense boundary and bounding box annotations, which can be prohibitively expensive. To bridge the gap, we investigate the weakly-supervised setting, where models learn from easily accessible video-language data without annotations. We identify that intra-sample spurious correlations among video-language components can be alleviated if the model captures the decomposed structures of video and language data. In this light, we propose a novel framework, namely WINNER, for hierarchical video-text understanding. WINNER first builds the language decomposition tree in a bottom-up manner, upon which the structural attention mechanism and top-down feature backtracking jointly build a multi-modal decomposition tree, permitting a hierarchical understanding of unstructured videos. The multi-modal decomposition tree serves as the basis for multi-hierarchy language-tube matching. A hierarchical contrastive learning objective is proposed to learn the multi-hierarchy correspondence and distinguishment with intra-sample and inter-sample video-text decomposition structures, achieving video-language decomposition structure alignment. Extensive experiments demonstrate the rationality of our design and its effectiveness beyond state-of-the-art weakly supervised methods, even some supervised methods.

1. Introduction

Spatio-temporal video grounding (STVG) is a fundamental task for video-language understanding [34, 41]. It aims to localize the spatio-temporal tube described by the language query from the untrimmed video. The essence of grounding lies in the semantic alignment of video and language components [34, 35, 41, 55]. To achieve such alignment, existing works fully exploit the fine-grained spatial-temporal annotations (*e.g.*, temporal boundaries, and spatial bounding boxes). For example, the spatial and temporal GCN modules in STGRN [55] require region-by-region frame-by-frame annotations for training convolutions.

Despite substantial research literature in this vein, the dense annotations (*i.e.*, video-by-video temporal boundary and frame-by-frame spatial bounding boxes) necessary for effective alignment require tremendous labor costs, which are, however, not routinely available. In addition, massive video-language data without spatial-temporal annotations are easily accessible but without exploitation. To reduce human labeling costs, we investigate the weakly supervised setting, where models learn to localize spatial-temporal tubes on easily accessible video-language data.

In the weakly-supervised setting, the absence of spatial-temporal annotations (*i.e.*, the exactly matched temporal boundaries and spatial bounding boxes) drives the video-language alignment problematic. Language semantics might get spuriously correlated with unmatched visual components in the untrimmed video. For example, when we inspect the sample shown in Figure 1 at either the video-sentence or the object-word hierarchy (*i.e.*, single-hierarchy understanding), the right-side `woman` can be hard to distinguish from the target `woman` to be localized. We identify that such spurious correlations could be alleviated if the model captures that the second `woman` is being pointed at when inspecting the sample at an intermediate video-language hierarchy. In this regard, we present a novel perspective for grounding-oriented video disambiguation and understanding under the weakly-supervised setting, *i.e.*, decomposing multiple aligned video-language hierarchies.

There are mainly two technical challenges to achieving such hierarchical alignment. On the one hand, the highly

*Equal Contribution.
†Corresponding Authors.

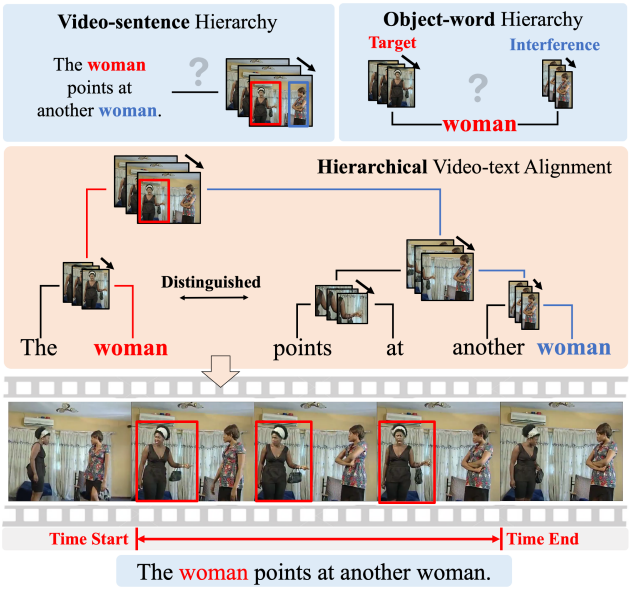

Figure 1. An illustration of how hierarchical video-language understanding alleviates spurious correlations in STVG.

unstructured nature of video data hinders its hierarchical understanding. Compared to syntactic parsing in language, where the elements are discrete tokens from a limited vocabulary, inferring the decomposed structure of continuous spatial-temporal visual data can be relatively challenging. On the other hand, hierarchical alignment requires simultaneously learning video-text correspondence and distinguishment at different hierarchies. However, in weakly supervised video grounding, the multi-hierarchy supervision signals (*e.g.*, paired and unpaired video-text components at different hierarchies) necessary for correspondence and distinguishment learning are rather lacking in raw data.

To address these challenges, we propose the **W**eakly-supervised h**I**erarchical decompositio**N** and alig**N**ment framework for spatio-t**E**mporal video g**R**ounding (**WINNER**). WINNER addresses *the first challenge* through language-structure guided video hierarchical understanding. In particular, WINNER first builds the language decomposition tree in a bottom-up manner. At each hierarchy, component-relevant visual cues are extracted and fused with the language decomposition tree through the proposed structural attention mechanism, resulting in a multi-modal decomposition structure. Such attention mechanism technically differs from existing ones by leveraging video-language information at adjacent hierarchies as context information, permitting the hierarchical understanding of unstructured video data. We introduce top-down feature backtracking over the multi-modal decomposition structure to ensure structural consistency and a text reconstruction objective to ensure the intra-sample correspondence between extracted visual cues and the multi-hierarchy language components.

Thanks to the multi-modal decomposition structure, the modality gap between video and text can be relatively narrowed in estimating the matching score between multi-hierarchy language components and potential spatial-temporal tube proposals. Upon the matching, we could build the language-grounded video decomposition tree, containing specific tubes as nodes, as revealed in Figure 1. A hierarchical contrastive learning objective is devised to recursively learn the desired correspondence and distinguishment over intra-sample and inter-sample video-text decomposition trees, thus addressing *the second challenge*.

We conduct experiments on two widely used datasets for the weakly-supervised video object grounding task. Experimental results show that the WINNER model greatly outperforms the state-of-the-arts, and could achieve comparable performance with some supervised methods. Extended experiments including the ablation study and the case study further demonstrate the rationality of the model involved.

Our contributions are summarized as follows:

- We research the spatio-temporal video grounding task under the challenging weakly supervised setting. We present a novel perspective, *i.e.*, hierarchical video-language decomposition and alignment, for alleviating spurious correlations brought by limited annotations.
- We propose a novel WINNER framework, which encapsulates the structural attention and top-down backtracking for multi-modal hierarchical understanding, and the multi-hierarchy intra-sample correspondence and inter-sample distinguishment learning.
- Experimental results demonstrate the rationality of our analysis and the effectiveness of WINNER.

2. Related Work

2.1. Spatio-temporal Video Grounding

The cross-modal visual grounding tasks, including image grounding [23, 36, 39, 48, 49, 57] and video grounding, have attracted the attention of more and more researchers. For the video grounding task, there are three fashions: temporal grounding, spatial grounding, and spatio-temporal grounding. The temporal video grounding [10, 15–17, 29, 37, 38, 46, 51] aims to detect the temporal clip described by the natural language query from the input video. Similarly, the spatial video grounding task [9, 20, 33] aims to localize the target video object described by the language sentence. Our focus is on the spatio-temporal video grounding task, a direction in which there is relatively little prior research.

The spatio-temporal video grounding task [11, 21, 22, 34, 40, 41, 54] is designed to detect the temporal boundary and the spatial object tube at the same time, according

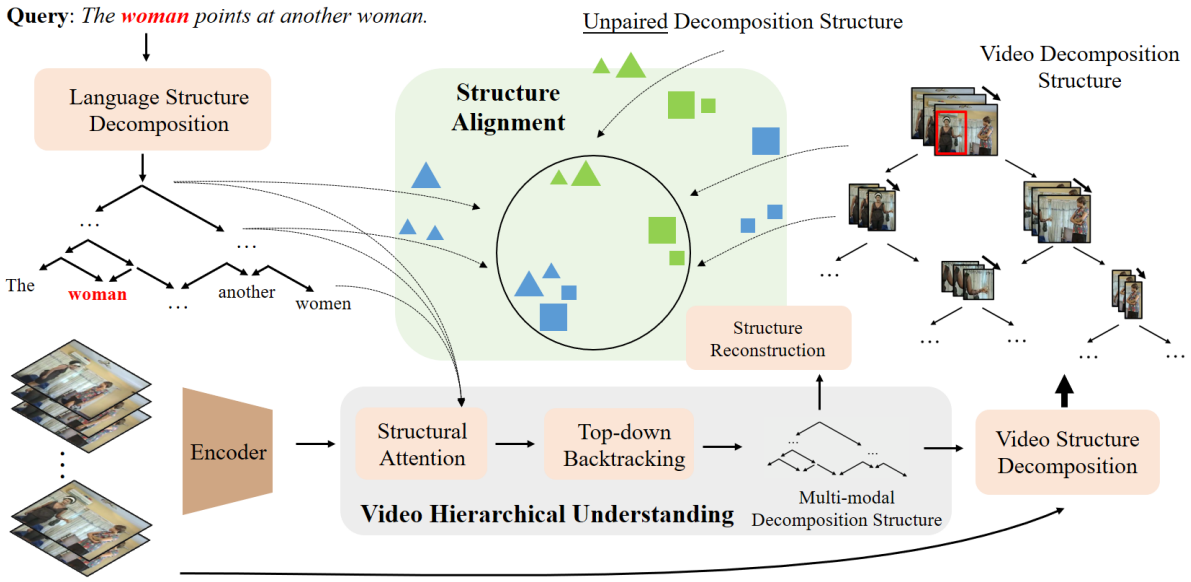

Figure 2. A schematic of the WINNER model. WINNER essentially encapsulates (1) the language structure decomposition; (2) structure-guided hierarchical video understanding, which attends to component-relevant visual cues with hierarchy consistency constraints; and (3) decomposition structure alignment, which decomposes the video into multi-hierarchy tubes through structure-tube matching, and learns the multi-hierarchy correspondence and distinguishment through hierarchical contrastive learning.

to the sentence. With the development of the neural network [6–8, 19, 28], it has become the mainstream for retrieval tasks [47], especially for the grounding task. [55] builds a spatio-temporal graph network to capture the relationships with temporal object dynamics and spatial tube dynamics. Some researchers try to improve performance through the attention mechanism [3, 13, 18, 45, 50]. [35] utilizes the visual transformer to extract cross-modal representation for visual object matching and temporal localization.

2.2. Weakly-supervised learning

Recently, weakly supervised learning has made significant progress in many areas [2, 4, 12, 42, 52], like the multi-modal field [5, 14, 24, 56]. Different from supervised learning [25–27, 43, 44], it achieves decent accuracy with very few annotations. [32] designs the deep learning model to complete the action segmentation task for the instructional videos with the semi-weakly supervised training way. [53] detects the language-described video clip with weakly supervised learning. [36] parses the video without annotations. [1] tries to learn the video grounding model in a weakly-supervised fashion without both the spatial bounding boxes and the temporal boundaries during the training process. However, previous works for the spatio-temporal grounding task simply focus on the feature of each frame and ignores the relation of adjacent frames. Different from them, we incorporate the unstructured context-dependent analysis into our model and use vision-language grammar induction to boost the performance of the spatio-temporal task in a weakly-supervised way.

3. Method

Spatio-temporal video grounding aims to localize the visual tubes $\mathcal{T}$ of a target object from an untrimmed video $\mathcal{V}$, where the target object corresponds to the subject of the input sentence $\mathcal{S}$. In the weakly supervised setting, models should learn from the video-language pairs $(\mathcal{S}, \mathcal{V})$, without ground-truth tube annotations $\mathcal{T}$. Apparently, this is a challenging setting where models might easily absorb spurious correlations among intra-sample video and language components. To bridge the gap, we propose the **W**eakly-supervised h**I**erarchical decompositio**N** and alig**N**ment framework for spatio-t**E**mporal video g**R**ounding (**WINNER**).

3.1. Model Overview

Our WINNER model is shown in Figure 2. We extract the region features from all video frames with pretrained Faster R-CNN [31] and obtain the corresponding tube feature via object tracking according to the regional coincidence degree, following existing techniques [54, 55]. With the tube features and the word features extracted by the pre-trained EMLo [30], the WINNER model mainly consists of three key processes: **(1) Cross-modal Hierarchical Understanding.** We regard words as leaf nodes and recursively build the language decomposition tree in a bottom-up layer-by-layer manner. Then, we propose the structural attention mechanism, which utilizes the language structure

Algorithm 1: The training process of the WINNER model.

Prepare: Initialize the language decomposition tree construction module $\mathcal{M}_{LSD}$ (language structure decomposition), the $\mathcal{M}_{SA}$ (structure attention), the $\mathcal{M}_{TDB}$ (top-down backtracking), and the $\mathcal{M}_{VSD}$ (video structure decomposition).

Input: The word features $\mathcal{F}_w$ of the language sentence $\mathcal{S}$; The tube features $\mathcal{F}_t$ of the input video $\mathcal{V}$.

Step 1: Build the language decomposition tree $\mathcal{T}_l$, $\mathcal{T}_l \leftarrow \mathcal{M}_{LSD}(\mathcal{F}_w)$;

Step 2: (1) Fuse the tube features $\mathcal{F}_t$ to get a multi-modal decomposition tree $\mathcal{T}_m$, $\mathcal{T}_m \leftarrow \mathcal{M}_{SA}(\mathcal{T}_l, \mathcal{F}_t)$;
(2) Update the tree $\mathcal{T}_m$ from top to down, $\mathcal{T}'_m \leftarrow \mathcal{M}_{TDB}(\mathcal{T}_m)$;
(3) Reconstruct the input language sentence $\mathcal{S}$ with the updated tree $\mathcal{T}'_m$ via the self-supervised learning;

Step 3: Realize the temporal localization based on the tree $\mathcal{T}'_m$;

Step 4: Generate the video tube tree $\mathcal{T}_t$, $\mathcal{T}_t \leftarrow \mathcal{M}_{VSD}(\mathcal{T}'_m, \mathcal{F}_t)$;

Step 5: Train the hierarchical alignment between the tree $\mathcal{T}'_m$ and the tree $\mathcal{T}_t$ with the contrastive learning.

as guidance and conducts hierarchical video understanding, resulting in a multi-modal decomposition structure. Finally, through top-down backtracking, we can ensure the hierarchical consistency of the multi-modal decomposition structure. **(2) Structure-guided Video Temporal Localization.** Under the guidance of temporal contrastive learning, the model learns to calculate the association between the multi-modal decomposition structure and different video clips, and select the most relevant video clip as the temporal localization result. Due to the limited paper space, we put this part in the appendix. **(3) Decomposition Structure Alignment.** We use the prediction of temporal localization in the second step to crop all the video tubes. Then, we train the model to find the most matching video tube for each node in the multi-modal decomposition tree trained by the hierarchical contrastive learning, thereby achieving cross-modal decomposition structure alignment. The video tube best matched with the decomposition tree node corresponding to the sentence subject is the predicted result. We summarize the whole training process in Alg 1.

3.2. Cross-modal Hierarchical Understanding

In order to achieve cross-modal hierarchical alignment, the hierarchical understanding of the input sentence $\mathcal{S}$ and the video $\mathcal{V}$ is a critical step. Technically, we first conduct language structure decomposition, followed by structure-guided video hierarchical understanding powered by structural attention and the top-down backtracking mechanisms. The main schematic is shown in Figure 3.

Language Structure Decomposition. A node $n_{i,j}$ in the language decomposition tree should contain the semantics from the i-th word to the j-th word in the sentence $\mathcal{S}$. In the beginning, there are $N_{\mathcal{S}}$ leaf nodes corresponding to $N_{\mathcal{S}}$ words in $\mathcal{S}$, upon which we build the language decomposition tree in a bottom-up manner by merging adjacent nodes. During the process, we denote the feature of node $n_{i,j}$ as $\mathbf{r}^B_{i,j}$, and its compatibility score as $c^B_{i,j}$. The compatibility score means how likely node $n_{i,j}$ will merge with other nodes. The i-th node feature $\mathbf{r}^B_{i,i}$ of the first layer is initialized with the i-th word embedding and $c^B_{i,i} = 0$. Then, all nodes of subsequent layers are generated based on their previous layer nodes. Specifically, for node $n_{i,j}$, the corresponding phrase (i, j) is divided into two parts with different methods. With a division (the phrase (i, k) and the phrase $(k + 1, j)$), we merge the features and the scores of the corresponding nodes, $n_{i,k}$ and $n_{k+1,j}$:

$$\mathbf{r}^B_{i,j,k} = MLP([\mathbf{r}^B_{i,k}, \mathbf{r}^B_{k+1,j}]), \tag{1}$$

$$c^B_{i,j,k} = (\mathbf{r}^B_{i,k})^T w(\mathbf{r}^B_{k+1,j}) + c^B_{i,k} + c^B_{k+1,j}, \tag{2}$$

where MLP represents the **M**ulti**L**ayer **P**erceptron, $[.]$ represents the feature concatenation, and w is a learnable parameter. All divisions are summarized into the feature $\mathbf{r}^B_{i,j}$ and the score $c^B_{i,j}$ of the tree node $n_{i,j}$:

$$\hat{c}^B_{i,j,k} = \underset{k}{softmax}(c^B_{i,j,k}), \tag{3}$$

$$\mathbf{r}^B_{i,j} = \sum_k \hat{c}^B_{i,j,k} * \mathbf{r}^B_{i,j,k}, \quad c^B_{i,j,k} = \sum_k \hat{c}^B_{i,j,k} * c^B_{i,j,k}, \tag{4}$$

Structure-guided Video Hierarchical Understanding. The structure-guided video hierarchical understanding contains two steps: (1) extracting multi-hierarchy visual cues relevant to components in the language decomposition tree and further transforming the language tree into a multi-modal one by feature fusion, powered by the *structural attention* mechanism; (2) recursively updating the multi-modal node representations through *top-down feature backtracking*, ensuring the hierarchical semantic consistency of the multi-modal decomposition tree. In the first step, assuming the node $n_{i,j}$ will be fused with the visual cues, we find all its related nodes, including its parent nodes $\{n_{t,k} : t \in [1, i); k \in (j, N_{\mathcal{S}}]\}$ and its child nodes $\{n_{t,k} : t, k \in [i, j]; t \leq k\}$. Then, we calculate the correlation between the language node features and the video tube features $\mathcal{F}^v = \{\mathbf{f}^v_m\}^{N_v}_{m=1}$, where N_v is the number of

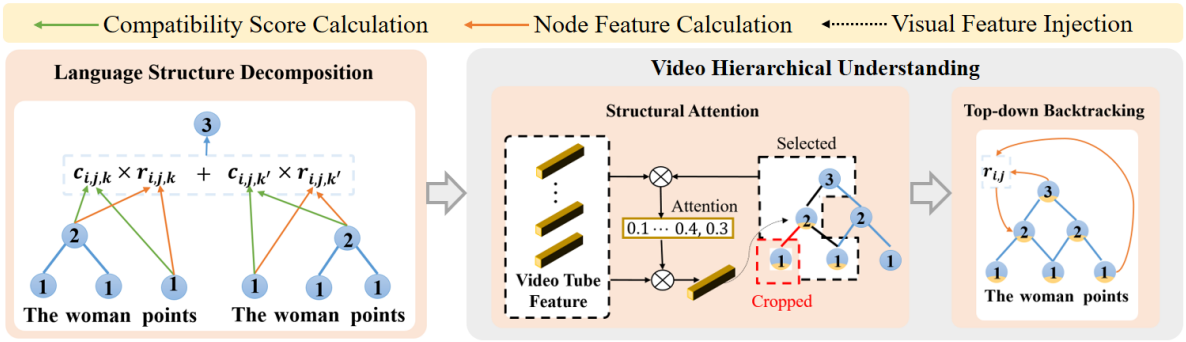

Figure 3. A schematic of the Cross-modal Hierarchical Understanding.

video tubes. Taking the video tube feature $\mathbf{f}_m^v$ and the node feature $\mathbf{r}_{t,k}^B$ as an example, the correlation $a_{t,k,m}$ is:

$$a_{t,k,m} = (\mathbf{f}_m^v)^T * \mathbf{r}_{t,k}^B. \tag{5}$$

We get the average attention value $\hat{a}_m$ for the video tube feature $\mathbf{f}_m^v$:

$$\hat{a}_m = \underset{(t,k)}{AVG}(a_{t,k,m}), \tag{6}$$

where AVG is the averaging function. Some tree nodes contain less meaningful information, such as the node corresponding to the word "the" or "at". The correlation between these nodes and the video information lacks reference value and might introduce interference. We judge whether the node $n_{t,k}$ is an interfering node, according to the similarity $s_{t,k}^c$ between its correlation values $a_{t,k} = \{a_{t,k,m}\}_{m=1}^{N_v}$ and the average correlation values $\hat{a} = \{\hat{a}_m\}_{m=1}^{N_v}$. The similarity $s_{t,k}^c$ is evaluated:

$$s_{t,k}^c = \frac{\sum_{m=1}^{N_v} a_{t,k,m} * \hat{a}_m}{\sqrt{\sum_{m=1}^{N_v}(a_{t,k,m})^2} * \sqrt{\sum_{m=1}^{N_v}(\hat{a}_m)^2}}. \tag{7}$$

If the similarity $s_{t',k'}^c$ for node $n_{t',k'}$ is less than the hyperparameter δ, we select out its correlation value $a_{t',k'} = \{a_{t',k',m}\}_{m=1}^{N_v}$. With the correlation values of the node $n_{i,j}$ and its parent (or child) nodes, we summarize them into the attention value $Att_{i,j,m}$ of node $n_{i,j}$ for video tube $\mathbf{f}_m^v$:

$$Att_{i,j,m} = \alpha * a_{i,j,m} + \beta * ((\sum\nolimits_{(t,k)} a_{t,k,m} - \sum\nolimits_{(t',k')} a_{t',k',m}) * \frac{1}{||j+t-i-k||_1}), \tag{8}$$

where α and β are learnable hyperparameters, $||.||_1$ is L1 normalization, and $a_{i,j,m}$ is the correlation value of the node $n_{i,j}$. $a_{t,k,m}$ and $a_{t',k',m}$ are the parent nodes or child nodes of the node $n_{i,j}$. Based on the attention values $\{Att_{i,j,m}\}_{m=1}^{N_v}$, we extract the visual cue $\mathbf{f}_{i,j}^v$ and inject it into the node feature $\mathbf{r}_{i,j}^B$ to get a new node feature $\widetilde{\mathbf{r}}_{i,j}^B$:

$$\mathbf{f}_{i,j}^v = \sum_m \underset{m}{softmax}(Att_{i,j,m}) * \mathbf{f}_m^v, \tag{9}$$

$$\widetilde{\mathbf{r}}_{i,j}^B = ||\mathbf{r}_{i,j}^B + \lambda * \mathbf{f}_{i,j}^v||_2, \tag{10}$$

where λ is a hyperparameter and $||.||_2$ is L2 normalization. After that, we repeat the bottom-up building process to update the next layer nodes with the new node feature $\widetilde{\mathbf{r}}_{i,j}^B$. In this way, we update the features of the decomposition tree nodes layer by layer.

In the second step, we update the features of all multi-modal decomposition tree nodes from top to down. For the tree node $n_{i,j}$, we represent the updated feature as $\mathbf{r}_{i,j}^U$ and the updated score as $c_{i,j}^U$. To generate them, we fuse the features and the scores of the nodes $n_{i,k}$ and $n_{k+1,j}$ for each $k \in [1,i) \cup (j, N_{\mathcal{S}}]$:

$$\mathbf{r}_{i,j,k}^U = MLP([\mathbf{r}_{i,k}^U, \widetilde{\mathbf{r}}_{j+1,k}^B]), \tag{11}$$

$$c_{i,j,k}^U = (\mathbf{r}_{i,k}^U)^T w(\widetilde{\mathbf{r}}_{j+1,k}^B) + c_{i,k}^U + c_{j+1,k}^B, \tag{12}$$

where w is a learnable parameter. We aggregate all features and scores corresponding to different k to get the updated feature $\mathbf{r}_{i,j}^U$ and the updated score $c_{i,j}^U$ of the node $n_{i,j}$:

$$\hat{c}_{i,j,k}^U = \underset{k}{softmax}(c_{i,j,k}^U), \tag{13}$$

$$\mathbf{r}_{i,j}^U = \sum_k \hat{c}_{i,j,k}^U * \mathbf{r}_{i,j,k}^U, \quad c_{i,j}^U = \sum_k \hat{c}_{i,j,k}^U * c_{i,j,k}^U. \tag{14}$$

We use the updated node features $\{\mathbf{r}_{i,i}^U\}_{i=1}^{N_{\mathcal{S}}}$ to reconstruct the input sentence $\mathcal{S}$ via self-supervised learning.

3.3. Decomposition Structure Alignment

In order to eliminate the ambiguity of single-level cross-modal matching, we adopt the method of multi-grained alignment to realize the spatio-temporal video grounding. Based on the multi-grained sentence features (detailed in Section 3.2), we structure and align the video features through cross-modal information retrieval.

Specifically, we crop the video clip out with the temporal boundary prediction in Section 3.3 and reconnect the object boxes of each frame extracted by Faster R-CNN as tubes.

The video tube features are represented as $\hat{\mathcal{F}}^v = \{\hat{\mathbf{f}}^v_m\}_{m=1}^{N_v}$, where N_v is the number of the video tubes. For each multi-modal decomposition tree node $n_{i,j}$, we calculate the similarity $s^s_{i,j,m}$ between its features and the video tube feature $\mathbf{f}^v_m$. Then for all the video tube features, we find the most similar one with the index m_{max}:

$$s^s_{i,j,m} = (\mathbf{f}^v_m)^T * (\tilde{\mathbf{r}}^B_{i,j} + \mathbf{r}^U_{i,j}), \tag{15}$$

$$m_{max} = \underset{m}{argmax}(s^s_{i,j,m}), \tag{16}$$

where the $argmax$ function is to find the index of the maximum value. After we find the best matching video tube for all multi-modal decomposition tree nodes, the video tube tree hierarchically aligned with the multi-modal decomposition tree is built up. With the aligned cross-modal information, the video tube corresponding to the sentence subject is viewed as the WINNER model prediction for the spatio-temporal video grounding task during the inference process.

At the training step, we adopt the spatial contrastive learning to train the hierarchical alignment. In detail, for the multi-modal decomposition tree node $n_{i,j}$, we get its maximum feature similarity $s^s(\mathcal{V}, n_{i,j})$ with the tubes of the video $\mathcal{V}$.

$$s^s(\mathcal{V}, n_{i,j}) = \underset{m}{max}(s^s_{i,j,m}). \tag{17}$$

For the i-th word, we maximize its similarity value $s^s(\mathcal{V}, n_{i,i})$, if this word comes from the sentence $\mathcal{S}$ paired with the video $\mathcal{V}$. Conversely, for the unpaired one $((\mathcal{V}', n_{i,i})$ or $(\mathcal{V}, n'_{i,i}))$, we minimize the similarity value $(s^s(\mathcal{V}', n_{i,i})$ or $s^s(\mathcal{V}, n'_{i,i}))$:

$$l_{word}(\mathcal{V}, n_{i,i}) = -log\frac{e^{s^s(\mathcal{V}, n_{i,i})}}{\sum_{\mathcal{V}'} e^{s^s(\mathcal{V}', n_{i,i})}}. \tag{18}$$

For the phrase (i, j), we need to consider how likely it exists in the input sentence $\mathcal{S}$:

$$d^s(\mathcal{V}, n_{i,j}) = s^s(\mathcal{V}, n_{i,j}) * \frac{c^B_{i,j} * c^U_{i,j}}{c^B_{1,N_\mathcal{S}}}. \tag{19}$$

Then, similar to the word level training, we maximize and minimize the similarity values of the paired data $((\mathcal{V}, n_{i,j}))$ and unpaired data $(s^s(\mathcal{V}', n_{i,j})$ or $s^s(\mathcal{V}, n'_{i,j}))$, respectively, for the phrase (i, j):

$$\begin{aligned} l_{phrase}(\mathcal{V}, n_{i,j}) = max(d^s(\mathcal{V}', n_{i,j}) - d^s(\mathcal{V}, n_{i,j}) \\ + \gamma, 0) + max(d^s(\mathcal{V}, n'_{i,j}) - d^s(\mathcal{V}, n_{i,j}) + \gamma, 0). \end{aligned} \tag{20}$$

The full loss function used to train the cross-modal hierarchical alignment is:

$$l_{full} = \mu * \sum_i l_{word}(\mathcal{V}, n_{i,i}) + \sum_{(i,j)} l_{phrase}(\mathcal{V}, n_{i,j}), \tag{21}$$

where μ is a hyperparameter.

4. Experiments

4.1. Experiment Settings

Datasets. To fully test our WINNER model, We adopt two widely used benchmarks for the spatio-temporal video grounding task, VidSTG [55] and HC-STVG [35]. **(1) VidSTG.** The VidSTG dataset is a large-scale dataset, which consists of $99,943$ sentences annotated on $6,770$ video clips. The natural language labels contain $55,135$ interrogative sentences and $44,808$ declarative sentences describing 79 types of objects. We follow the official split. **(2) HC-STVG.** The HC-STVG dataset is sampled from the movies and annotated with human-centered. There are $5,660$ video-query pairs in the dataset, in which 57.2% of video clips contain more than 3 people. We refer to the VidSTG dataset for preprocessing. The official dataset split is followed during the model testing process.

Implement Detail. We experiment on a Linux server to train our WINNER model. Our model is implemented using the pytorch framework with many other tools, such as numpy, torchvision, and so on. In the process of model implementation, we set $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.5$. The λ and the μ are both set as 1.0. During the training process, the learning rate is $1e-5$, and the batch size is 64. We train the WINNER model with the maximum epoch set as 20.

Evaluation Criteria. We follow the evaluation protocol [34, 35] and use vIoU@R and m_vIoU as the evaluation criteria. Specifically, the S_U and the S_I are the union and the intersection of the predicted and ground-truth frames, respectively. The vIoU is calculated by $vIoU = \frac{1}{|S_U|}\sum_{t\in S_I} IoU(r^t, \hat{r}^t)$. In it, the r^t and $\hat{r}^t$ are the bounding boxes of the model prediction and the ground truth, respectively. The vIoU@R is the data proportion of $vIoU > R$, and the m_IoU is the average of all samples' IoU values.

4.2. Performance Comparison

Comparison with the State-of-the-arts. We adopt several weakly supervised video grounding state-of-the-arts as the baselines for comparison: AWGU [1] and Vis-Ctx [33]. The performance comparison results are shown in Table 1. From the table, we have the following findings: (1) We observe that the performance of all models on the VidSTG declarative sentence grounding is better than their performance on the VidSTG interrogative sentence grounding. This is probably because, in the absence of the subjects in interrogative language queries, the spurious correlation between the interference (another word in the sentence with the same noun as the hidden subject) and the target object could be potentially stronger. (2) The baselines, including AWGU and Vis-Ctx, perform worse than our WINNER model. These methods generally neglect the spurious correlations among video and language components

Table 1. Comparison results between WINNER and the state-of-the-arts on the VidSTG and the HC-STVG datasets. Larger scores indicate better performance. We color each row as the **best** and **second best**. Acronym notations of each model can be found in Section 4.1.

Methods	VidSTG (Declarative Sentence)			VidSTG(Interrogative Sentence)			HC-STVG		
	m_vIoU	*vIoU@0.3*	*vIoU@0.5*	*m_vIoU*	*vIoU@0.3*	*vIoU@0.5*	*m_vIoU*	*vIoU@0.3*	*vIoU@0.5*
AWGU	8.96	**7.86**	3.10	8.57	6.84	2.88	8.20	4.48	0.78
Vis-Ctx	**9.34**	7.32	**3.34**	**8.69**	**7.18**	**2.91**	**9.76**	**6.81**	**1.03**
WINNER (Ours)	**11.61**	**14.12**	**7.40**	**10.23**	**11.96**	**5.46**	**14.20**	**17.24**	**6.12**

Table 2. Comparison results between WINNER and fully supervised baselines. Notably, WINNER is trained under the weakly supervised setting.

Methods	*m_vIoU*	*vIoU@0.3*	*vIoU@0.5*
Declarative Sentence Grounding			
GroundeR	9.78	11.04	4.09
STPR	10.40	12.38	4.27
WSSTG_T	11.36	14.63	5.91
WSSTG_L	14.45	18.00	7.89
STGRN	19.75	25.77	14.60
WINNER (Ours)	11.61	14.12	7.40
Interrogative Sentence Grounding			
GroundeR	9.32	11.39	3.24
STPR	9.98	11.74	4.36
WSSTG_T	10.65	13.90	5.32
WSSTG_L	13.36	17.39	7.06
STGRN	18.32	21.10	12.83
WINNER (Ours)	10.23	11.96	5.46

under the weakly supervised setting by aligning the sentence and the video as a whole. (3) WINNER consistently outperforms baseline methods across different metrics and datasets. We contribute these merits to the decomposition of video-text structures, upon which the multi-hierarchy intra-sample correspondence and inter-sample distinguishment could be achieved, alleviating spurious associations between multi-modal components.

Comparison with Fully Supervised Methods. We are interested in how our weakly supervised WINNER model performs compared to the supervised spatio-temporal video grounding methods. In this light, we follow [55] to choose several widely used baselines for comparison. The experiment results on the VidSTG dataset are shown in Table 2. From this table, we surprisingly find that our weakly supervised WINNER model outperforms several supervised baselines (GroundeR and STPR) and achieves comparable performance to some others (*e.g.*, WSSTG_T). These results further demonstrate the effectiveness of WINNER.

Table 3. Ablation study of WINNER on the VidSTG dataset. **DSA** is the **D**ecomposition **S**tructure **A**lignment, and **SVHU** represents the **S**tructure-guided **V**ideo **H**ierarchical **U**nderstanding.

DSA	SVHU	*m_vIoU*	*vIoU@0.3*	*vIoU@0.5*
Declarative Sentence Grounding				
		7.42	7.61	2.60
✓		9.97	11.65	5.27
	✓	10.89	13.08	6.40
✓	✓	11.61	14.12	7.40
Interrogative Sentence Grounding				
		7.07	7.33	2.78
✓		8.79	9.66	4.08
	✓	9.65	10.91	5.13
✓	✓	10.23	11.96	5.46

4.3. In-depth Analysis

Ablation Study. We further evaluate the contribution of the key modules in our WINNER model. In detail, we surgically remove the **D**ecomposition **S**tructure **A**lignment (**DSA**) and the **S**tructure-guided **V**ideo **H**ierarchical **U**nderstanding (**SVHU**) from the WINNER model and get different architectures. Removing DSA would result in a total loss of sentence-guided video content detection, which is necessary for grounding. As such, instead of removing DSA, we replace it with contrastive learning at the word-tube level without multiple hierarchies. Ablation study results on the VidSTG dataset are shown in Table 3. According to the results, we have several observations: (1) Removing either SVHU or DSA would lead to a significant performance drop, which demonstrates the effectiveness of the two modules in achieving hierarchical video-text under-

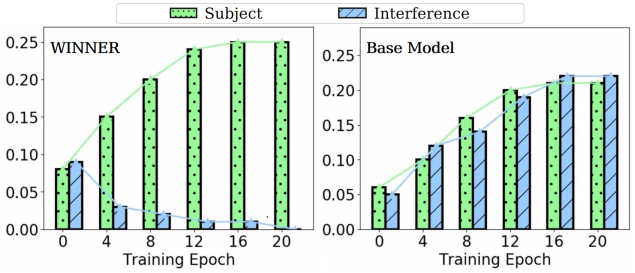

Figure 4. The correlation score between the subject or the inference (same noun as the subject) in the sentence and the target video tube changes with the training epochs for our WINNER model and the ablation base model.

standing and alignment, respectively. (2) When SVHU and DSA are used in combination, the model performs better than either alone. It reflects that the two modules can enhance each other, and both of them are indispensable.

Spurious Correlation Alleviation. We are interested in whether the WINNER model successfully improves the alleviation of spurious correlations. In this light, we present the averaged correct correlation scores between the sentence subjects and the ground truth video tubes, as well as the averaged spurious correlation scores between the interference nouns (same word in the sentence as the subject but with different context, such as `A woman points at another` *`woman`*) and the ground truth video tubes. How to calculate the correlation score is in the appendix. The averaged correct/spurious correlation scores of the WINNER model and the base model, as illustrated in the ablation study part, across different training epochs are shown in Figure 4. According to the results, we can find that WINNER could successfully learn to alleviate spurious correlations and improve correct correlations. In contrast, the base model without hierarchical video-language understanding still suffers from the spurious correlations between the interference noun and the ground truth video tube. These results further demonstrate the rationality of our analysis of spurious correlations and the effectiveness of multi-hierarchy alignment in the weakly supervised setting.

Case Study. Figure 5 presents the spatial-temporal grounding results of three cases predicted by the WINNER model. We observe that WINNER precisely locates the video clips corresponding to the input language query. In the weakly supervised setting, models would easily absorb intra-sample spurious correlations among language and video components (*e.g.*, `another adult in black` and the target video tube in the first case.). WINNER firstly analyzes the language structure, which figures out the relationships (*e.g.*, `leans on`) between language components, upon which WINNER further decomposes the video structure and conducts multi-hierarchy alignment. Such a hierarchical understanding of video-language data

Figure 5. Case study of the spatio-temporal grounding results for the WINNER model.

intuitively helps to alleviate potential spurious correlations and contributes to the effectiveness under the weakly supervised setting.

5. Conclusion

In this paper, we introduce a novel perspective, hierarchical video language decomposition and alignment for the spatio-temporal video grounding task to alleviate the spurious correlation. Then, we propose the WINNER framework, which consists of the cross-modal hierarchical understanding and the decomposition structure alignment. Extensive experiments demonstrate the rationality of our analysis and the effectiveness of WINNER.

6. Acknowledgments

Our work is supported by the National Natural Science Foundation of China under Grant No. 62222211, No.61836002, and No.62072397. In addition, our research is funded in part by the Program of Zhejiang Province Science and Technology (No.2022C01044), NSFC (No.62037001), and the Starry Night Science Fund at Shanghai Institute for Advanced Study of Zhejiang University.

References

[1] Junwen Chen, Wentao Bao, and Yu Kong. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In *ACM MM*, 2020. 3, 6

[2] Andrew Drozdov, Patrick Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In *NAACL*, 2019. 3

[3] Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. Dependency parsing as mrc-based span-span prediction. In *ACL*, 2022. 3

[4] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In *CVPR*, 2021. 3

[5] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, 2021. 3

[6] Wei Ji, Xi Li, Lina Wei, Fei Wu, and Yueting Zhuang. Context-aware graph label propagation network for saliency detection. *IEEE TIP*, 2020. 3

[7] Wei Ji, Xi Li, Fei Wu, Zhijie Pan, and Yueting Zhuang. Human-centric clothing segmentation via deformable semantic locality-preserving network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 3

[8] Wei Ji, Xi Li, Yueting Zhuang, Omar El Farouk Bourahla, Yixin Ji, Shihao Li, and Jiabao Cui. Semantic locality-aware deformable network for clothing segmentation. In *IJCAI*, 2018. 3

[9] Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. Vidvrd 2021: The third grand challenge on video relation detection. In *ACM MM*, 2021. 2

[10] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-Seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *CVPR*, 2023. 2

[11] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *arXiv preprint arXiv:2209.13306*, 2022. 2

[12] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *CVPR*, 2022. 3

[13] Ming Kong, Qing Guo, Shuowen Zhou, Mengze Li, Kun Kuang, Zhengxing Huang, Fei Wu, Xiaohong Chen, and Qiang Zhu. Attribute-aware interpretation learning for thyroid ultrasound diagnosis. *Artificial Intelligence in Medicine*, 2022. 3

[14] Sangmin Lee, Hyung-Il Kim, and Yong Man Ro. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *CVPR*, 2022. 3

[15] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu, and Yueting Zhuang. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *arXiv preprint arXiv:2301.09071*, 2023. 2

[16] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, 2022. 2

[17] Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, and Yueting Zhuang. Dilated context integrated network with cross-modal consensus for temporal emotion localization in videos. In *ACM MM*, 2022. 2

[18] Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu. Multi-task attribute-fusion model for fine-grained image recognition. In *Optoelectronic Imaging and Multimedia Technology VII*, 2020. 3

[19] Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. Ib-m: A flexible framework to align an interpretable model and a black-box model. In *BIBM*, 2020. 3

[20] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *ACL*, 2022. 2

[21] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *ACM MM*, 2022. 2

[22] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Stvgformer: Spatio-temporal video grounding with static-dynamic cross-modal understanding. In *Proceedings of the 4th on Person in Context Workshop*, 2022. 2

[23] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *CVPR*, 2021. 2

[24] Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Wenping Wang. Multi-view depth estimation using epipolar spatio-temporal networks. In *CVPR*, 2021. 3

[25] Zheqi Lv, Zhengyu Chen, Shengyu Zhang, Kun Kuang, Wenqiao Zhang, Mengze Li, Beng Chin Ooi, and Fei Wu. Ideal: Toward high-efficiency device-cloud collaborative and dynamic recommendation system. *arXiv preprint arXiv:2302.07335*, 2023. 3

[26] Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, and Fei Wu. Duet: A tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*, 2023. 3

[27] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 3

[28] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 3

[29] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv*, 2020. 2

[30] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 3

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 3

[32] Yuhan Shen and Ehsan Elhamifar. Semi-weakly-supervised learning of complex actions from instructional task videos. In *CVPR*, 2022. 3

[33] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*, 2019. 2, 6

[34] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021. 1, 2, 6

[35] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*. 1, 3, 6

[36] Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *ICLR*, 2021. 2, 3

[37] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *ACM MM*, 2020. 2

[38] 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

[39] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding by contrastive knowledge distillation. In *CVPR*, 2021. 2

[40] Zeyu Xiong, Daizong Liu, and Pan Zhou. Gaussian kernel-based cross modal network for spatio-temporal video grounding. In *ICIP*, 2022. 2

[41] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 1, 2

[42] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *CVPR*, 2021. 3

[43] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 3

[44] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[45] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *ACM MM*, 2020. 3

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

[47] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *SIGIR*, 2021. 3

[48] Wenqiao Zhang, Jiannan Guo, Mengze Li, Haochen Shi, Shengyu Zhang, Juncheng Li, Siliang Tang, and Yueting Zhuang. Boss: Bottom-up cross-modal semantic composition with hybrid counterfactual training for robust content-based image retrieval. *arXiv preprint arXiv:2207.04211*, 2022. 2

[49] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *AAAI*, 2021. 2

[50] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE TMM*, 2019. 3

[51] Wenqiao Zhang, Xin Eric Wang, Siliang Tang, Haizhou Shi, Haochen Shi, Jun Xiao, Yueting Zhuang, and William Yang Wang. Relational graph learning for grounded video description generation. In *ACM MM*, 2020. 2

[52] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *CVPR*, 2022. 3

[53] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, 2020. 3

[54] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Nicholas Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. *arXiv*, 2020. 2, 3

[55] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020. 1, 3, 6, 7

[56] 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. 3

[57] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *CVPR*, 2018. 2