# Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Spatio-Temporal video grounding (STVG) focuses on retrieving the spatio-temporal tube of a specific object depicted by a free-form textual expression. Existing approaches mainly treat this complicated task as a parallel frame-grounding problem and thus suffer from two types of inconsistency drawbacks: *feature alignment inconsistency* and *prediction inconsistency*. In this paper, we present an end-to-end one-stage framework, termed Spatio-Temporal Consistency-Aware Transformer (STCAT), to alleviate these issues. Specially, we introduce a novel multi-modal template as the global objective to address this task, which explicitly constricts the grounding region and associates the predictions among all video frames. Moreover, to generate the above template under sufficient video-textual perception, an encoder-decoder architecture is proposed for effective global context modeling. Thanks to these critical designs, STCAT enjoys more consistent cross-modal feature alignment and tube prediction without reliance on any pre-trained object detectors. Extensive experiments show that our method outperforms previous state-of-the-arts with clear margins on two challenging video benchmarks (VidSTG and HC-STVG), illustrating the superiority of the proposed framework to better understanding the association between vision and natural language.

## 1 Introduction

Visual grounding is a prominent and fundamental task in the multi-modal understanding field. It aims to localize a region from the visual content specified by a given natural language query. As a bridge to connect vision and language, this topic has drawn increasing research attention over the past few years [13, 11, 46, 18, 16]. Early works mostly focused on static images and have achieved remarkable progress, while visual grounding in videos has not been adequately explored yet. Recently, spatio-temporal video grounding (STVG) was introduced in [49], which is a compound task, requiring both spatial and temporal localization. Formally, given an untrimmed video and a textual description of an object, this task aims at producing a spatio-temporal tube (*i.e.*, a sequence of bounding boxes) for the queried target object (See Figure 1). Compared with previous grounding tasks in images, STVG is substantially more challenging as it delivers high demands for distinguishing subtle spatio-temporal status of instance in videos based on the query sentence. Thus, how to effectively align the textual semantics and time-varying visual appearance is especially critical for settling this task.

The majority of current approaches [39, 35, 4, 49, 48] streamline the spatio-temporal video grounding as a two-stage pipeline, where the object proposals or tubelets are firstly generated by a pre-trained object detector (*e.g.*, Faster RCNN [26]) and then all the candidates are ranked according to the similarity with the query. However, this simple design leads to an inevitable deficiency: the grounding performance is heavily limited by the quality of proposals, especially when a huge domain gap exists between the pre-trained object categories and free-form text. Lately, some approaches [30, 34] try to

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

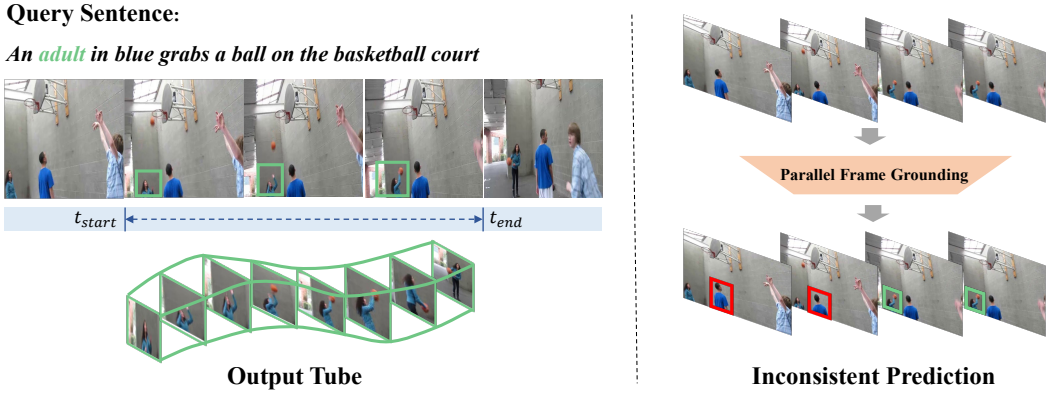


Figure 1: An illustration of spatio-temporal video grounding task (left) and the inconsistency prediction in previous approaches (right). This task is particularly challenging and requires exquisite reasoning about linguistic semantics and the global video context. Existing methods simply treat STVG as a parallel image grounding problem, leading to inconsistent predictions (highlighted in red).

address these issues by leveraging a one-stage framework without any off-the-shelf object detectors. They directly produce the bounding box at each frame as well as predict the starting and ending boundary based on the fused cross-modal features.

Though promising results have been achieved, most of these one-stage approaches have the following drawbacks that need to be concerned. (1) **Feature alignment inconsistency**. Global context modeling is of great significance for STVG scenario. For example, given the query of "An adult in blue grabs a ball on the basketball court" as in Figure 1, the target "adult" cannot be determined without a perception of the whole video content, since there are two adults in blue but only one of them eventually grabs the ball. Accordingly, the cross-modal feature alignment should be performed between the whole video and the querying sentence. However, existing methods [49, 30, 34] only attend the cross-modal fusion with local short-term video context. Even in the very recent method [41], the alignment is still considered at frame-level. (2) **Prediction inconsistency**. Previous approaches simply treat STVG as a parallel frame-level grounding problem. As shown in Figure 1, they separately generate a bounding box with respect to the query on each single frame, yet neglecting the consistency among frames. It is worth noting that the goal of STVG task is to localize a single target instance specified by the textual description. Although the appearance of the target instance across all video frames may change due to camera motion, scene dynamics, etc., they should have the identical semantics (*e.g.*, the guy who grabs the ball). Therefore, conducting grounding independently on each frame regardless of semantic consistency over time will cause the prediction mismatch.

In this paper, we present an effective one-stage transformer-fashion encoder-decoder based [36] framework, dubbed as Spatio-Temporal Consistency-Aware Transformer (STCAT) to cope with the aforementioned issues in the STVG task. Building upon the prevalent detection network DETR [2], STCAT introduces a novel video-level multi-modal template to modulate the grounding procedure. This template serves as a global target that guides and correlates the predictions among each individual frame to maintain temporal consistency. Specially, it is formulated as a learnable query as in [2], which is comprised of a content term shared by all frames to encode the semantics and a position term characterized by per frame. Such design allows the final grounding results to be determined based on the specific frame content while simultaneously considering consistency. Moreover, in order to adaptively obtain the above-mentioned query by considering the interactions between visual and textual information, we propose a cross-modal spatio-temporal encoder to fuse the global video context and textual description. The query in STCAT is then rendered based on the encoded features and fed into a decoder to aggregate the multi-modal representation. During the decoding procedure, the query belonging to each frame is dynamically updated layer by layer. Finally, the grounded spatio-temporal tubelet is produced by different prediction heads.

The technical contributions of this work can be summarized as follows. (1) We propose a novel multi-modal template mechanism to mitigate the prediction inconsistency issue in the STVG task. In doing so, the grounding among all video frames can be correlated to yield a more precise spatio-temporal prediction. (2) We design a transformer-based architecture, named as STCAT, to fully exploit the global video context and thus perform a better cross-modal alignment. Coupling with the proposed

template mechanism, our STCAT can directly ground the spatio-temporal tubes without reliance on any pre-trained object detectors. (3) Comprehensive experiments conducted on two challenging video benchmarks (VisSTG [49] and HC-STVG [34]) further demonstrate that our method obtained new state-of-the-art performance compared to other approaches.

## 2 Related Work

**Visual Grounding in Images/Videos.** Visual grounding in images/videos is a popular task, which is aimed to localize the object of interest in an image/video based on a text description. In the image grouding scenario, most existing methods [19, 45, 21, 37, 42] firstly utilize a pre-trained object detector to get some object proposals. While some recent works like [16, 44, 24, 43] propose one-stage frameworks that alleviate the reliance on pre-trained detectors. For example, in [16] Liao *et al.* proposed to extract cross-modal representations and localize the target objects by an anchor-free object detection method. Yang *et al.* tried to generate text-conditional visual features by sub-queries in [43], which further promote the one-stage method.

For the video grounding task, which can be categorized into temporal grounding and spatio-temporal grounding. The former requires to localizes a temporal clip in the video by given referring sentence. While the spatio-temporal video grounding lies at the intersection of spatial and temporal localization. Most approaches [4, 34, 39, 49] for STVG also rely on pre-extracted tube or object proposals. Recently, STVGBert in [30] propose a one-stage approach that extend the VilBERT [23] to settle this task. However, all these methods suffer from the drawbacks mentioned in Section 1. More recently, a concurrent work in [41] also proposed a one-stage framework that utilizes DETR [2] architecture. However, they only consider the multi-modal interaction at frame-level and still lack a well design for settling prediction inconsistency in STVG, thus achieve inferior performance than ours.

**Vision-language modeling.** Vision-language modeling tasks such as visual question answering, image captioning, and image-text retrieval attract lots of researchers to explore recently. Due to the simplicity and success of Transformers in the natural language processing field, many works [33, 14, 23, 31] borrow this architecture to align the context information between images and sentences. Others like [32, 51, 6, 9, 15, 40] further extend transformers into video-text tasks. But most of them rely either on pre-extracted object features, or spatially pooled features, that ignored detailed spatial information within each frame, and are not capable to handle the challenging STVG task.

**Transformer based detection.** Object detection is a traditional task in computer vision field, early methods [26, 25] usually rely on CNN encoders and region proposal or regression modules to tackle the various objects in images. As Carion *et al*. [2] first introduce Transformers into this task, a lot of transformer-based detection methods [52, 1, 50] have sprung up recently. Most of them follow the encoder-decoder paradigm to transform the visual features into precision object bounding box, we borrow this effective architecture to tackle the STVG task in this paper.

## 3 The Proposed Approach

In this section, we briefly present the formal definition of the STVG task and an overview of our proposed framework in Section 3.1. Then we elaborate on the main components of STCAT, including the feature extractor (Section 3.2) and the consistency-aware transformer (Section 3.3). Finally, the training and inference details are introduced in Section 3.4.

### 3.1 Overview

Given an untrimmed video $\mathbf{V} = \{v_t\}_{t=1}^{T}$ consisting of $T$ consecutive frames and a query textual description $\mathbf{S} = \{s_n\}_{n=1}^{N}$ depicting a target object existing in $\mathbf{V}$. The goal of the STVG task is to localize a spatio-temporal tube $\mathbf{B} = \{b_t\}_{t=t_s}^{t_e}$ corresponding to the semantics of given textual query, where $b_t$ represents a bounding box in the $t$-th frame, $t_s$ and $t_e$ specify the starting and ending boundary of the retrieved object tube respectively. STVG is an extremely challenging task, which requires not only the spatial interaction with the textual modality for detecting a bounding box at frame level, but also the long-range temporal relation modeling at video level for determining the start and end timestamps of query-related video segments.

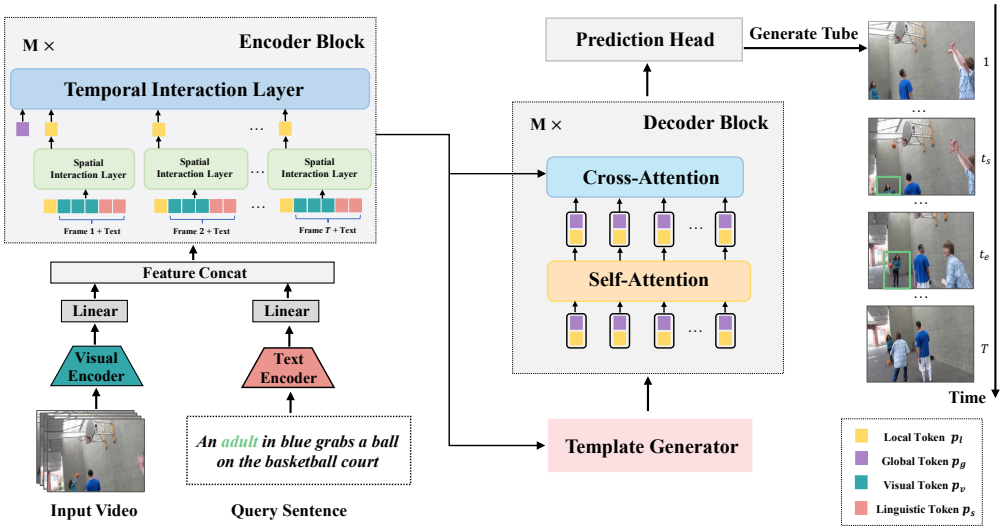


Figure 2: The Architecture of the proposed Spatio-Temporal Consistency-Aware Transformer (STCAT). Given an input video and query sentence pair, it firstly leverages a visual and a linguistic encoder to extract features for each modality. The extracted features are fed into a encoder to perform cross-model interaction. Then a generator yields the multi-modal templates which is responsible for guiding the decoder. The retrieved object tube is finally generated based on the decoded features via a prediction head.

To mitigate the aforementioned feature alignment and prediction inconsistency drawbacks, an effective framework named STCAT is developed in this paper. As illustrated in Figure 2, the proposed model firstly utilizes two feature extractors to obtain both visual and textual features from the video frames and querying sentence, respectively. It then models the video-text interactions through a well-designed spatio-temporal cross-modal encoders, which introduces a global learnable token for the whole video to encode target object semantics, and a local one for each individual frame to represent the frame-specific appearance. These tokens are further leveraged to produce a template for target object by a template generator. Finally, the yielded template is treated as a query (Like the one in DETR [2]) and fed into a decoder to aggregate features and predict the retrieved spatio-temporal tube.

## 3.2 Feature Extractor

**Visual Encoder.** We start by adopting a vision backbone (*e.g.*, ResNet [10]) to extract the visual features for each frame $v_t$ individually in video $\mathbf{V} = \{v_t\}_{t=1}^{T}$. The obtained 2D feature map of per frame is then flattened, bringing a visual feature sequence $\mathcal{F}_v = \{f_t^v\}_{t=1}^{T}$. Each $f_t^v \in \mathcal{R}^{N_v \times C_v}$ serves as a compact representation for $v_t$, where $N_v = H \times W$ and $C_v$ is the visual feature channel.

**Linguistic Encoder.** For the querying sentence $\mathbf{S} = \{s_n\}_{n=1}^{N_S}$ with $N$ words, we leverage a pre-trained linguistic embeding encoder, (*e.g.*,BERT model [5]), to encode the linguistic representation. The textual features of input sentence query is denoted as $\mathcal{F}_s = \{f_i^s\}_{i=1}^{N}$, where $f_i^s \in \mathcal{R}^{C_s}$ and $C_s$ is textual feature channel.

## 3.3 Consistency-Aware Transformer

The detailed spatio-temporal video grounding procedure is executed by leveraging four crucial components: a cross-modal transformer encoder, a template generator, a query-guided transformer decoder and two parallel prediction heads.

### 3.3.1 Multi-modal Encoding

The proposed encoder is obliged to fully exploit the cross-modal interaction between the video and text, correlating their corresponding semantics in a fine-grained manner. Existing one-stage video grounding approaches only restrict this interaction either in individual frames [41] or short video snippets [30]. This design leads to an inconsistent cross-modal feature alignment as the query sentence depicts a long-term evolutional event in videos.

To this end, the developed encoder aims at performing a more consistent feature alignment between two modalities. Specially, given the input features $\mathcal{F}_v$ and $\mathcal{F}_s$, a projection layer is firstly applied to embed them into the same channel dimension $C$. We denote the projected visual embedding as $p_v = \{p_{v_t}\}_{t=1}^{T}$, where $p_{v_t} \in \mathcal{R}^{N_v \times C}$ and linguistic embedding as $p_s \in \mathcal{R}^{N_s \times C}$. Besides, we introduce a learnable token denoted as $p_g \in \mathcal{R}^{\mathcal{C}}$ to integrate the global video context during encoding, and $T$ analogous ones dubbed as $p_l = \{p_l^t \in \mathcal{R}^{C}\}_{t=1}^{T}$ to attend the local context within individual frame. The proposed encoder reads all above-described embedding tokens as its input and has $M$ stacked encoder blocks. Each block consists of a spatial and a temporal interaction layer. Both adopt the transformer-style encoder structure [36].

**Spatial Interaction Layer.** The goal of this layer is to conduct the intra- and inter-modality relation modeling spatially for each local frame. In detail, the input $x_t$ for each interaction layer on the $t$-th frame can be formulated as:

$$x_t = [p_l^t, \underbrace{p_{v_t}^1, p_{v_t}^2, ..., p_{v_t}^{N_v}}_{\text{visual tokens } p_{v_t}}, \underbrace{p_s^1, p_s^2, ..., p_s^{N_s}}_{\text{linguistic tokens } p_{s_t}}]. \tag{1}$$

The joint input sequence $x_t$ is then fed into one transformer encoder layer to yield the contextualized visual-text representations for each frame respectively. It is worth noting that the state of frame-specific token $p_l^t$ is enriched by both visual and linguistic context within frame $v_t$.

**Temporal Interaction Layer.** The aforementioned spatial interaction layer only attends the local context information at frame level but lacks the modeling of global context across the whole video. To address this, our temporal interaction layer further models the interactions temporally between local frames. Similarly, the input $x$ belonging to this layer is fed into another transformer encoder layer, which can be formulated as:

$$x_g = [p_g, \underbrace{p_l^1, p_l^2, ..., p_l^T}_{\text{frame tokens } p_l}]. \tag{2}$$

To retain the temporal relative information, we also add a positional encoding to $x_g$. Through this temporal layer, the local frame tokens $p_l = \{p_l^t\}_{t=1}^{T}$ attend the whole video content and the global token also aggregates the global video-text context.

After the above encoding procedure, we harvest the contextualized multi-modal features $F_{vl} \in \mathcal{R}^{T \times (N_v + N_s) \times C}$, the global embedding $p_g$ for the whole video and the local embedding $p_l = \{p_l^t\}_{t=1}^{T}$ with respect to all $T$ video frames.

### 3.3.2 Template Generation

Unlike previous works that ground each frame separately without considering the consistency, we design a template-based mechanism to correlate and restrict the predictions across all video frames. Concretely, the developed template is comprised of a content term and a position term. The content term $\boldsymbol{q_c}$ is shared by all frames to hold identical semantics depicted by the query sentence, while the position term $\boldsymbol{q_p} = \{\boldsymbol{q_p^t}\}_{i=1}^{T}$ is characterized by per frame according to its appearance. Both of them are yielded by a template generator based on the encoded local and global tokens.

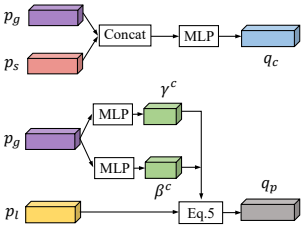


Figure 3: The illustration of procedure for yielding content term $q_c$ and position term $q_p$ in the template generator.

As illustrated in Figure 3, the content term $\boldsymbol{q_c} \in \mathcal{R}^{C}$ is generated by jointly considering the global visual-linguistic context:

$$\boldsymbol{q_c} = \boldsymbol{W}_c[p_g; p_s] + \boldsymbol{b_c}, \tag{3}$$

where $p_s$ is a sentence-level feature obtained by the average pooling of textual tokens produced by encoder, $[;]$ is a concatenation operation. Both of $\boldsymbol{W}_c$ and $\boldsymbol{b_c}$ are learnable parameters. As for the position term $\boldsymbol{q_p} = \{\boldsymbol{q_p^t}\}_{i=1}^{T}$, where each $\boldsymbol{q_p^t}$ is a 4-dimensional vector dubbed as $(x_t, y_t, w_t, h_t)$, which serves as a reference anchor to restrict the grounding region on each individual frame $v_t$. Based on considerations of both global semantics consistency and local characteristic specificity are critical for $\boldsymbol{q_p}$, we generate it via depending both local frame embedding $p_l$ and global embedding $p_g$ (See Figure 3). In detail, inspired by [47, 43], we

firstly leverage $p_g$ to modulate the local embedding $p_l = \{p_l^t\}_{t=1}^T$. It is fulfilled by two linear layers to generate two modulation vectors $\gamma^c \in \mathbf{R}^C$ and $\beta^c \in \mathbf{R}^C$ based on $p_g$:

$$\gamma^c = \tanh\left(W_\gamma p_g + b_\gamma\right), \quad \beta^c = \tanh\left(W_\beta p_g + b_\beta\right), \tag{4}$$

where $\boldsymbol{W_\gamma}$, $\boldsymbol{b_\gamma}$, $\boldsymbol{W_\beta}$ and $\boldsymbol{b_\beta}$ are learnable parameters. Finally, the position term $\boldsymbol{q_p^t}$ for each specific frame $v_t$ is obtained by modulating $p_l^t$:

$$\boldsymbol{q_p^t} = \text{Sigmoid}\left(f_p\left(\gamma^c \odot p_l^t + \beta^c\right)\right), \tag{5}$$

where $f_p$ is a learnable mapping layer: $\mathcal{R}^C \rightarrow \mathcal{R}^4$.

#### 3.3.3 Query-Guided Decoding

The objective of our decoder is to convert the multi-modal representations $F_{vl}$ obtained from the encoder to the target object tube based on the generated template. Inspired by recent query-based detectors [38, 20, 8], we formulate the above template $(\boldsymbol{q_c}, \boldsymbol{q_p})$ as object queries to guide the overall decoding procedure, where $\boldsymbol{q_c}$ serves as content queries to probe target object patterns in encoded embedding and $\boldsymbol{q_p}$ acts as positional queries to restrict the possible attending regions [20].

In detail, we denote $\{Q_t\}_{t=1}^T$ as the object query for each individual frame $v_t$. The initial query for feeding the decoder is formulated by $Q_t = [C_t; P_t]$ and we have:

$$C_t = \boldsymbol{q_c}, \quad P_t = \text{Linear}(\text{PE}(\boldsymbol{q_p^t})), \tag{6}$$

where PE means the sinusoidal position encoding to $\boldsymbol{q_p^t} = (x_t, y_t, w_t, h_t)$. The decoder adopts a standard architecture in line with DETR [2]. Each decoder block includes a self-attention module for feature updating across all video frames, and a cross-attention module for feature probing within the corresponding frame $v_t$. The content query $C_t$ is thus enriched by stacking $M$ decoder layers, resulting in the final contextualized representation used for generating object tubes. Following the previous practice [52, 38], during decoding, the position query $P_t$ are also updated in each layer via predicting a relative offset with respect to $\boldsymbol{q_p^t}$ by a shared regression head described below.

**Prediction Head.** Finally, the refined content query is utilized for simultaneously predicting the temporal boundary and bounding boxes on each frame. In detail, a regression head implemented as a 3-layer MLP is adopted to perform bounding box coordinates prediction. The output of this prediction head is a 4-dimensional coordinate offset $(\Delta x_t, \Delta y_t, \Delta w_t, \Delta h_t)$ and the final predicted box $\hat{b_t} \in [0, 1]^4$ in frame $v_t$ is obtained by:

$$\hat{b_t} = (x_t + \Delta x_t, y_t + \Delta y_t, w_t + \Delta w_t, h_t + \Delta h_t). \tag{7}$$

The temporal boundary $[\hat{t_s}, \hat{t_e}]$ is fulfilled by predicting the start and end probabilities for each frame with a similar MLP prediction head.

### 3.4 Training and Inference

**Training.** At the training stage, we send a batch of video-sentence pairs to the proposed model. Each pair has a ground-truth bounding box sequence $\mathbf{B} = \{b_t\}_{t=t_s}^{t_e}$ and the corresponding start and end timestamps, $t_s$ and $t_e$. For spatial localization, we involve the box prediction loss $\mathcal{L}_{bbox}$ as follows:

$$\mathcal{L}_{bbox} = \lambda_{L_1}\mathcal{L}_{L_1}(\hat{B}, B) + \lambda_{\text{giou}}\mathcal{L}_{giou}(\hat{B}, B), \tag{8}$$

where $\mathcal{L}_{L_1}$ and $\mathcal{L}_{\text{giou}}$ are the smooth $L_1$ loss and generalized IoU loss [27] on the bounding boxes respectively. Note that $\mathcal{L}_{bbox}$ only considers predictions in $[t_s, t_e]$. As for temporal localization, we follow [28, 30] to generate two 1-dimensional heatmaps $\pi_s, \pi_e \in \mathcal{R}^T$ for starting and ending positions. And the temporal prediction loss is termed as $\mathcal{L}_{\text{temp}} = \mathcal{L}_s(\hat{\pi_s}, \pi_s) + \mathcal{L}_e(\hat{\pi_e}, \pi_e)$, where $\mathcal{L}_s$ and $\mathcal{L}_e$ are the focal loss [17]. To encourage a more accurate temporal prediction, we also add an auxiliary head to predict whether a frame belongs to the ground-truth segment. It is supervised by a binary cross-entropy loss term $\mathcal{L}_{\text{seg}}$. The total training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{bbox} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}. \tag{9}$$

**Inference.** During inference, our framework produces the bounding boxes and the starting and ending prediction probabilities for all frames. The start and end times of the output tube, $\hat{t_s}$ and $\hat{t_e}$, are computed by selecting the frames with the joint starting and ending score with invalid combinations $\hat{t_s} \geq \hat{t_e}$ removed. The bounding boxes within $[\hat{t_s}, \hat{t_e}]$ form the final retrieved tube.

# 4 Experiments

## 4.1 Datasets and Metrics

**Datasets.** To evaluate the proposed method, we follow the previous work [30] and adopt two large video grounding benchmarks: VidSTG [49] and HC-STVG [34]. Both of the two datasets are annotated with spatio-temporal tubes corresponding to text queries. **VidSTG** contains in total 6,924 untrimmed videos, which are clipped into 44,808 video-triplet pairs with 99,943 sentences describing 80 types of objects. We follow the pioneer work [49] to split all videos into three parts, 5,563 for training, 618 for validation and 743 for testing, which paired with 80,684, 8,956 and 10,303 distinct sentences respectively. **HC-STVG** contains 5,660 untrimmed videos in multi-person scenes, each annotated with one expression related to human attributes or actions. To keep the same settings as previous works, this dataset is divided into the training set and the testing set with 4,500 and 1,160 video-sentence pairs, respectively.

**Evaluation Metrics.** For fair comparison, we follow [49] and select **m_vIoU**, **m_tIoU** and **vIoU@R** as the evaluation criteria. Note that vIoU $= \frac{1}{|S_u|}\sum_{t\in S_i}\text{IoU}(\hat{b}_t, b_t)$, where $S_i$ and $S_u$ are the intersection and union between the predicted tubes and the ground-truth tubes, respectively. And $\text{IoU}(\hat{b_t}, b_t)$ describes the IoU score between the detected bounding box $\hat{b}_t$ and the ground-truth bounding box $b_t$ at frame $t$. While the m_vIoU score is defined as the average vIoU score over all testing videos. As for vIoU@R, which is the ratio of samples whose vIoU > R in the testing subset. Besides that, we also use m_tIoU (the mean of tIoU) to evaluate the temporal localization performance only, where tIoU = $\frac{|S_i|}{|S_u|}$.

## 4.2 Implementation Details

In line with the previous methods, We adopt the ResNet-101 [10] as the visual encoder to extract the visual features and RoBERTa [22] as the linguistic encoder. The whole framweork is end-to-end optimized for directly predicting the target object tubes. The visual linguistic encoder and our cross-modal encoder are initialized with pre-trained weights provided in [12], while the decoder is trained from scratch. Following [49, 41], the input videos are sampled at a frame rate of 5 fps and we further uniformly down-sample the videos with more than 200 sampled frames. Besides, due to the appearance similarity between adjacent frames, the cross-modal interaction is only conducted for every $K$-th frame for computation efficiency during encoding. We empirically set the hyper-parameters in this paper with $M = 6, C = 256, K = 3$ and the loss weights as $\lambda_{L_1} = 5, \lambda_{\text{giou}} = 2, \lambda_{\text{temp}} = 10$ and $\lambda_{\text{seg}} = 2$. The batch size and base learning rate are set to be 32 and $1e^{-4}$, respectively. More details can be found in our supplemental material.

## 4.3 Performance Comparison

To fully demonstrate the superiority of the proposed model, we compare our approach to all previous spatio-temporal video grounding methods on VidSTG and HC-STVG datasets. Specially, we introduce three types of competitors: (a) **Factorized**: These approaches factorize STVG task as performing spatial and temporal grounding separately. They firstly utilize temporal visual grounding techniques like TALL [7] and L-Net [3] to predict the temporal boundary of target object and then perform spatial visual grounding with methods like GroundeR [29], STPR [39] and WSSTG [4] to generate the bounding boxes. (b) **Two-Stage**. The two-stage approaches like STGRN [49], STGVT [34] and OMRN [48] accomplish spatio-temporal grounding via firstly generating box proposals in each frame by means of a pre-trained object detector and then selecting the best matches from these candidates. (c) **One-Stage**. The one-stage methods like STVGBert [30] and a very recent concurrent work TubeDETR [41] tackle the STVG task through a unified architecture that directly grounds the final spatio-temporal object tube from the given video-sentence pair.

Table 1: Performance comparisons of the state-of-the-art on the VidSTG test set (%).

| Methods | Declarative Sentences | | | | Interrogative Sentences | | | |
|---|---|---|---|---|---|---|---|---|
| | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
| ***Factorized:*** | | | | | | | | |
| GroundeR [29]+TALL [7] | | 9.78 | 11.04 | 4.09 | | 9.32 | 11.39 | 3.24 |
| STPR [39]+TALL [7] | 34.63 | 10.40 | 12.38 | 4.27 | 33.73 | 9.98 | 11.74 | 4.36 |
| WSSTG [4]+TALL [7] | | 11.36 | 14.63 | 5.91 | | 10.65 | 13.90 | 5.32 |
| GroundeR [29]+L-Net [3] | | 11.89 | 15.32 | 5.45 | | 11.05 | 14.28 | 5.11 |
| STPR [39]+L-Net [3] | 40.86 | 12.93 | 16.27 | 5.68 | 39.79 | 11.94 | 14.73 | 5.27 |
| WSSTG [4]+L-Net [3] | | 14.45 | 18.00 | 7.89 | | 13.36 | 17.39 | 7.06 |
| ***Two-Stage:*** | | | | | | | | |
| STGRN [49] | 48.47 | 19.75 | 25.77 | 14.60 | 46.98 | 18.32 | 21.10 | 12.83 |
| STGVT [34] | - | 21.62 | 29.80 | 18.94 | - | - | - | - |
| OMRN [48] | **50.73** | 23.11 | 32.61 | 16.42 | 49.19 | 20.63 | 28.35 | 14.11 |
| ***One-Stage:*** | | | | | | | | |
| STVGBert [30] | - | 23.97 | 30.91 | 18.39 | - | 22.51 | 25.97 | 15.95 |
| TubeDETR [41] | 48.10 | 30.40 | 42.50 | 28.20 | 46.90 | 25.70 | 35.70 | 23.20 |
| STCAT (Ours) | 50.58 | **32.14** | **46.31** | **30.56** | **49.40** | **27.40** | **39.10** | **25.29** |

Table 2: Performance comparisons of the state-of-the-art on the HC-STVG test set (%).

| Methods | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|
| ***Two-Stage:*** | | | | |
| STGVT [34] | - | 18.15 | 26.81 | 9.48 |
| ***One-Stage:*** | | | | |
| STVGBert [30] | - | 20.42 | 29.37 | 11.31 |
| TubeDETR [41] | 43.70 | 32.40 | 49.80 | 23.50 |
| STCAT (Ours) | **48.74** | **35.35** | **58.79** | **31.98** |

The complete performance comparisons between our model and the aforementioned approaches are presented on Tables 1 and 2 for two video benchmarks respectively. Overall, our STCAT achieves the state-of-the-art grounding performance in terms of all evaluation metrics (**m_vIoU**, **m_tIoU** and **vIoU@R**) for both two complex benchmarks, except a little bit lower accuracy in **m_tIoU** on VidSTG (v.s. OMRN [48]), demonstrating the superiority of the proposed model. Although slight preponderance in **m_tIoU**, OMRN leverages an off-the-shelf detector, while our model is end-to-end trainable. From the shown results, we can further obtain the following observations. 1) Separating the STVG task into individual spatial and temporal grounding problems achieved unsatisfactory performance in comparison to tacking them simultaneously (worse performance than both one-stage and two-stage). 2) Compared with conventional two-stage methods, the removal of object detectors brings a significant performance boost. It is mainly due to that the one-stage methods are end-to-end trainable and break through the restriction of the domain gap imposed by the pre-trained detectors. 3) Our proposed method outperforms all existing one-stage methods on two datasets. Although the very recent method [41] also applies the DETR architecture to STVG task, it only performs cross-modal interaction at the frame level in the encoder and implicitly models the temporal correlation through the self-attention layer in the origin DETR architecture. In contrast, our framework considers the global video context and maintain a template mechanism as queries for explicitly modeling consistency. All these novel designs contribute to the empirical superiority to [41] by large margins. Especially for the HC-STVG benchmark (*e.g.*, 31.98% v.s. 23.50% in terms of **vIoU@0.5**), since it contains more longer query sentence to describe a complicated scene change within movies, delivering a higher demand for modeling subtle spatio-temporal clues..

### 4.4 Ablation Study

**Effect of the template mechanism.** To demonstrate the validation of the proposed global and local template mechanism, we design three variants of our models : 1) Remove the global template in decoder and initialize the content query with a zero vector, while remain the local template for generating position query. 2) Remove the local template in decoder and leverage a learnable position query while remain the global template as content query. 3) Remove both two types of template. The quantitative ablation results are shown in Table 3a. One can observe that cancelling either local or global template mechanism will weaken the grounding performance. Besides, the local template contributes more to the whole performance (+2.0% m_vIoU v.s. +0.7% m_vIoU). This is basically intuitive since the local template is specified by frame characteristic. In summary, both two types of template are benefitial for the STVG task.

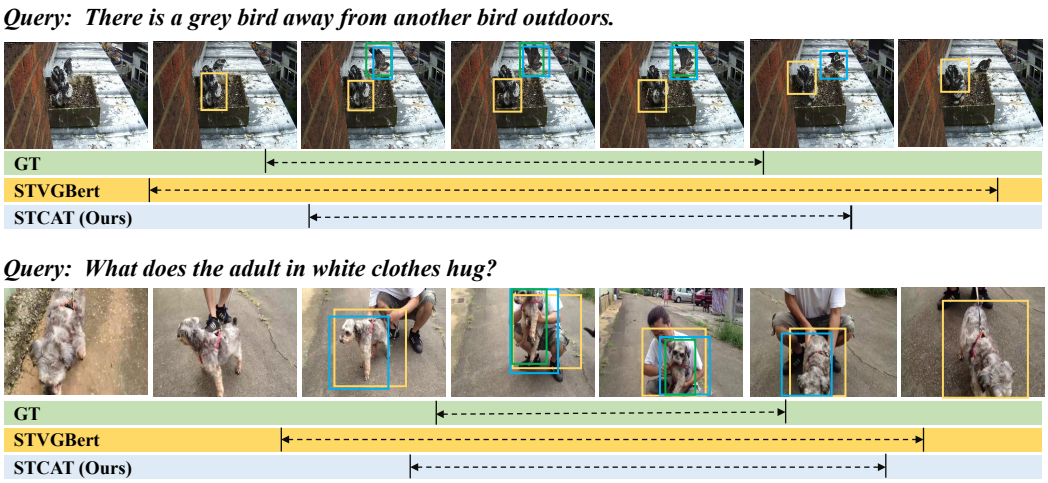


Figure 4: Some illustration examples of the spatio-temporal grounding results produced by STVGBert [30] (yellow) and our model (blue), compared with ground truth (green) on VidSTG benchmark.

| Local | Global | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|
| | | 29.48 | 41.80 | 27.57 |
| ✓ | | 31.46 | 44.90 | 30.09 |
| | ✓ | 30.17 | 42.16 | 28.52 |
| ✓ | ✓ | **32.14** | **46.31** | **30.56** |

(a) Ablations for the local and global templates.

| Setting | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|
| w/o Temp | 30.92 | 44.60 | 29.24 |
| w/ Temp | **32.14** | **46.31** | **30.56** |

(b) Ablations for the temporal interaction layer.

Table 3: The ablation studies on VidSTG benchmark (declarative sentences) for demonstrating the effectiveness of different components in our proposed STCAT.

**Effect of the temporal interaction layer.** The temporal interaction layer in our cross-modal encoder is obliged to model the global context across the whole video, which is crucial for conducting a consistent feature alignment for encoder and maintaining a video-level objective for the decoder. We thus explore the function of temporal interaction layer via eliminating it from each encoder block (denoted as w/o temp). The detailed ablation results are shown in Table 3b. From the presented comparison, we can observe a distinct performance drop without this layer, which further validates the effectiveness of our proposed temporal interaction layer.

### 4.5 Qualitative Analysis

In this section, we illustrate some examples in figure 4 to compare our predictions with the one-stage method STVGBert [30] on VidSTG dataset to qualitatively validate the effectiveness of proposed approach. As shown in Figure 4, our model is able to predict meaningful and accurate spatio-temporal tubes than STVGBert. Specially, for the first query, the ground truth tube should contain the top-right bird in the given video specified by “away from". However, due to the lack of long-range context modeling, the STVGBert retrieved the unmatched results. In contrast, our STCAT attended the whole video content and thus can precisely localize the desired instance. More visualization examples are provided in supplementary material.

## 5 Conclusion

In this paper, we develop an effective one-stage consistency-aware transformer network (STCAT) for tackling the spatio-temporal video grounding task. To mitigate the feature alignment and prediction inconsistency in previous approaches, our proposed STCAT introduces a novel template mechanism to modulate the grounding procedure, which serves as a global target that guides and correlates the predictions among each individual frame to maintain temporal consistency. In addition, The extensive experimental results further demonstrate the superiority of the proposed network on the STVG task.

# References


[1] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182, 2019.

[4] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020.

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[8] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. *arXiv preprint arXiv:2203.16507*, 2022.

[9] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Co-operative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124, 2017.

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[14] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[15] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[16] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[18] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019.

[19] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017.

[20] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[21] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[24] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[27] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[28] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2464–2473, 2020.

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[30] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021.

[31] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[32] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[33] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[34] 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*, 2021.

[35] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019.

[38] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based object detection. *arXiv preprint arXiv:2109.07107*, 2021.

[39] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1453–1462, 2017.

[40] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[41] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. *arXiv preprint arXiv:2203.16434*, 2022.

[42] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019.

[43] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[44] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019.

[45] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[46] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[47] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1069–1075, 2021.

[49] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.

[50] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020.

[51] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Presented in our supplementary material.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]