# CoSTA: End-to-End Comprehensive Space-Time Entanglement for Spatio-Temporal Video Grounding

**Yaoyuan Liang*[1], Xiao Liang*[1], Yansong Tang†[1], Zhao Yang[2], Ziran Li[3], Jingang Wang[3], Wenbo Ding[1], Shao-Lun Huang[1]**

[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling,
Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]University of Oxford,
[3]Meituan Inc.

## Abstract

This paper studies the spatio-temporal video grounding task, which aims to localize a spatio-temporal tube in an untrimmed video based on the given text description of an event. Existing one-stage approaches suffer from insufficient space-time interaction in two aspects: i) less precise prediction of event temporal boundaries, and ii) inconsistency in object prediction for the same event across adjacent frames. To address these issues, we propose a framework of Comprehensive Space-Time entAnglement (CoSTA) to densely entangle space-time multi-modal features for spatio-temporal localization. Specifically, we propose a space-time collaborative encoder to extract comprehensive video features and leverage Transformer to perform spatio-temporal multi-modal understanding. Our entangled decoder couples temporal boundary prediction and spatial localization via an entangled query, boasting an enhanced ability to capture object-event relationships. We conduct extensive experiments on the challenging benchmarks of HC-STVG and VidSTG, where CoSTA outperforms existing state-of-the-art methods, demonstrating its effectiveness for this task.

## Introduction

Spatio-temporal video grounding (STVG) (Huang et al. 2018; Su, Yu, and Xu 2021; Yang et al. 2022a; Jin et al. 2022) is the task of localizing natural language in untrimmed videos, where a linguistic expression describes an event happening to an object, and the objective is to predict a spatio-temporal tube (*i.e.*, a sequence of bounding boxes) that localizes that object. Different from conventional spatial (Liu et al. 2019a,b; Yang et al. 2020; Deng et al. 2021; Li and Sigal 2021) and temporal grounding tasks (Anne Hendricks et al. 2017; Zhang et al. 2019b; Chen et al. 2020; Zhang et al. 2020a, 2021), which focus specifically on either spatial or temporal dimensions, STVG requires joint space-time localization based on the given linguistic query. As a result, it poses a more difficult challenge in respect to multi-modal learning—the modeling and exploitation of spatio-temporal dependencies in visual features.

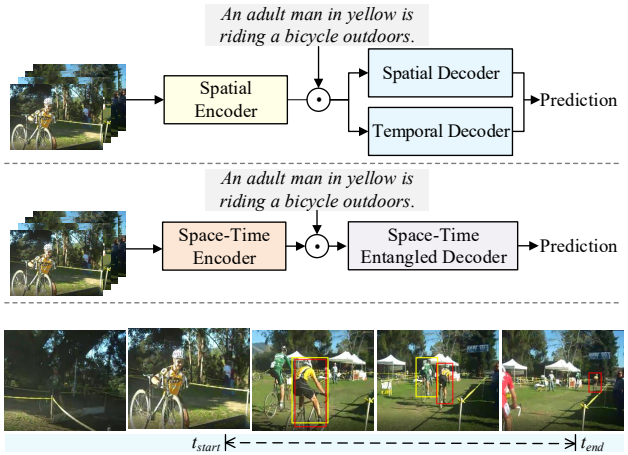


Figure 1: The task of spatio-temporal video grounding takes an untrimmed video and a text description as inputs, and predicts a spatio-temporal tube that corresponds to the event and the object described. **(a)** The previous state-of-the-art method (Jin et al. 2022) employs a spatial encoder (*e.g.*, 2D CNN) to extract frame-wise video features and decodes the results in a parallel way. **(b)** Our CoSTA encodes both spatial and temporal video features and utilizes an entangled decoder to produce the final predictions. **(c)** The spatio-temporal tubes predicted by our method (in red) and by that of (Jin et al. 2022) (in yellow), highlighting our method's temporal consistency in object grounding by comparison.

An increasing amount of attention has been drawn to the STVG task over the past few years. In recent literature, two-stage and one-stage models are the two mainstream approaches to capturing multi-modal correspondences in the video and text inputs. The two-stage paradigm (Yamaguchi et al. 2017; Tang et al. 2021; Zhang et al. 2020c) is to first extract object proposals (Zhang et al. 2020c) or tube proposals (Tang et al. 2021) by leveraging a pre-trained object or tube detector. The proposals are then ranked based on cross-modal similarity measures, from which one proposal is selected as the prediction. Such approaches heavily rely on the pre-trained tube extractor and cannot recover from proposal failures. Meanwhile, the success of end-to-end Transformer-

*These authors contributed equally.
†Corresponding author.

based object detection systems (*e.g.*, (Carion et al. 2020; Meng et al. 2021; Liu et al. 2022)) has inspired several of the recent developments of one-stage approaches (Yang et al. 2022a; Jin et al. 2022). Typically, they interpret the video as a sequence of individual frames and employ a Transformer encoder to perform frame-wise multi-modal feature encoding. Then tube decoding (the process of predicting a tube from the multi-modal features) is factorized into two subtasks (see Fig. 1a): temporal boundary prediction and bounding box prediction, which are achieved by parallel prediction heads (Yang et al. 2022a) or a pair of decoupled decoders (Jin et al. 2022) based on query-to-feature similarity. The two types of boundaries are composed in a post-processing step to generate the final spatio-temporal tube.

Although great progress has been made, the exploration for effective means of interaction between spatial and temporal information (residing in both visual and linguistic data) is far from sufficient in the existing paradigms. Specifically, language information is fused with individual frames and multi-modal encoding does not exploit temporal information. As a consequence, these approaches are less effective at learning features that are discriminative of the temporal boundaries of events, which is critical to obtaining accurate tube predictions in this task. This phenomenon is highlighted in Fig. 2, where we report the $vIoU$ and $sIoU$ metrics[1] of the state-of-the-art method (Yang et al. 2022a; Jin et al. 2022) in comparison with those of our method. Obtaining a good $vIoU$ critically depends on the accurate prediction of temporal boundaries, while doing well in $sIoU$ mainly concerns with bounding box predictions over the spatial dimensions. It can be seen that the existing approaches perform well as measured by $sIoU$ but struggle to obtain a high $vIoU$, and by comparison, our proposed method achieves a significantly higher $vIoU$. Moreover, at the parallel tube decoding stage of the existing approaches, the absence of mutual awareness between spatial location of the object and time boundary of the event potentially leads to an inconsistency in spatial grounding over consecutive frames (depicted in Fig. 1c).

To address the issues above, we propose Comprehensive Space-Time entAnglement (CoSTA), an end-to-end framework for the STVG task. The "comprehensiveness" is reflected in the following way: We conduct space-time interaction in the video encoding, the multi-modal fusion, and the tube decoding stages, whereas the existing approach (Jin et al. 2022) only performs it during multi-modal fusion. As shown in Fig. 1b, CoSTA is built upon an encoder-decoder architecture, in which the encoder fully integrates spatial and temporal features in a video, with the purpose of facilitating motion-aware multi-modal reasoning, and the decoder makes bounding box predictions and temporal boundary predictions in a connected way that exploits the dependency between the duration of the event and the target object. More concretely, we capture the visual features via the proposed space-time collaborative encoder under a hierarchical scheme, and then utilize a Transformer encoder (Arnab et al. 2021; Bertasius, Wang, and Torresani 2021; Jin et al. 2022) to perform joint spatio-temporal multi-modal alignment. Afterwards, our decoder bridges temporal boundary prediction and spatial localization via the proposed "entangled query," which enforces the model to localize the target object according to the correlation between the duration of the event and the object's spatial appearance. This framework enables fuller spatio-temporal interaction, and as a result, yields more accurate grounding in both space and time. We conduct extensive experiments and obtain new state-of-the-art results on two challenging STVG benchmarks, HC-STVG (Tang et al. 2021) and VidSTG (Zhang et al. 2020c), which demonstrates the effectiveness of the proposed method.

[1]The definitions of $sIoU$ and $vIoU$ are detailed in *Experiments/Datasets and Metrics/Evaluation metrics* section. Intuitively, the ground-truth time boundaries are known when calculating $sIoU$, but unknown when calculating $vIoU$.

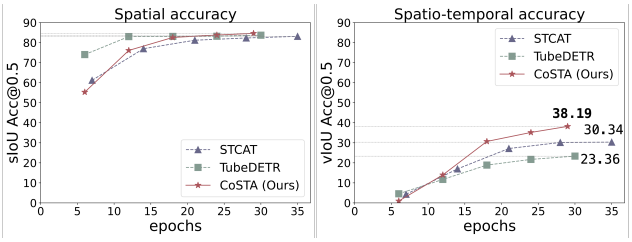


Figure 2: Comparisons between our method and other state-of-the-art methods (Yang et al. 2022a; Jin et al. 2022) in sIoU and vIoU. Left: Existing methods have comparable performance with ours on spatial-only grounding. Right: Our method obtains significantly higher $vIoU$ while having a faster convergence rate.

## Related Work

**Spatio-temporal video grounding (STVG)**. This task aims to retrieve a spatio-temporal tube from an untrimmed video corresponding to a given sentence query, which involves performing both temporal video grounding and frame-wise spatial grounding. Annotating both temporal boundaries and frame-wise bounding boxes in a video is a complex and labor-intensive task. To address this challenge, earlier STVG methods focus on solving this problem under weakly-supervised settings (Huang et al. 2018; Shi et al. 2019; Chen et al. 2019; Chen, Bao, and Kong 2020; Tan et al. 2021). Concurrently, some supervised methods (Yamaguchi et al. 2017; Tang et al. 2021; Zhang et al. 2020c) employ a pre-trained tube extractor to generate tube proposals as potential candidates. These proposals are ranked based on their similarity with the sentence query, and the best-matching proposal is selected as the final prediction.

Without pretrained proposal extractors, STGVBert (Su, Yu, and Xu 2021) extends ViLBERT (Lu et al. 2019) to the video grounding task and simultaneously models both spatial and temporal interactions in an end-to-end way. Inspired by MDETR (Kamath et al. 2021) in image grounding, TubeDETR (Yang et al. 2022a) develops a one-stage Transformer encoder-decoder model for STVG, with slow-fast multi-modal encoding and parallel space-time decoding. Most recently, STCAT (Jin et al. 2022) employs a Trans-

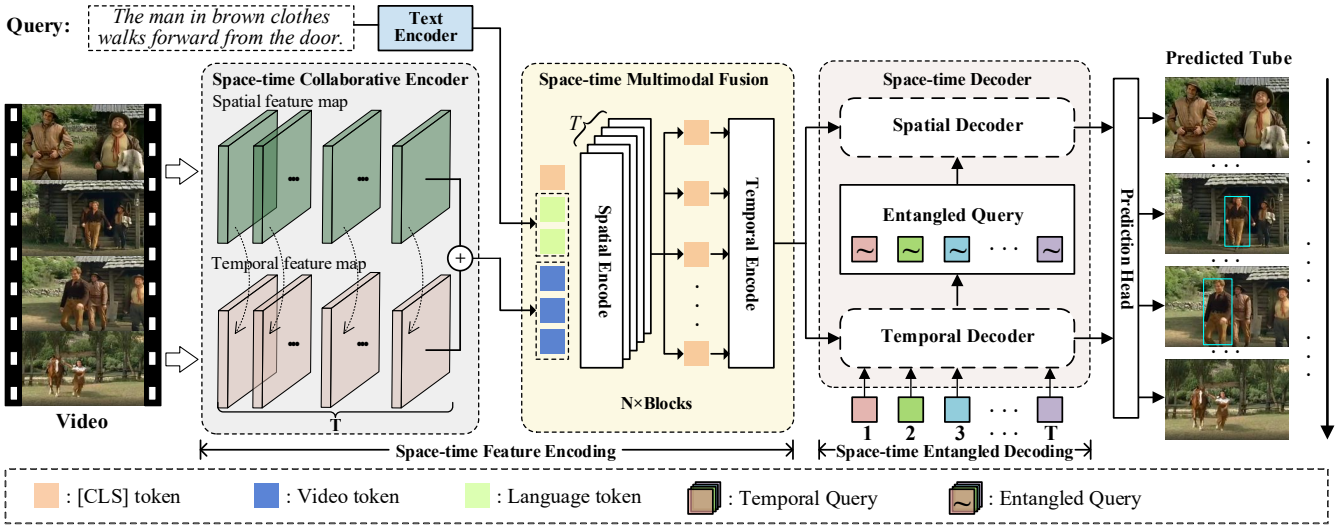


Figure 3: Overview of the proposed CoSTA. Given an untrimmed video and a sentence description, our model first employs a proposed space-time video encoder (illustrated in Fig. 4) and a text encoder (*e.g.*, RoBERTa (Liu et al. 2019c)) to extract unimodal features respectively. Then the extracted features are concatenated and sent to the multi-modal encoder, which is composed of $N$ spatio-temporal encoding layers. Finally, our proposed space-time entangled decoder (illustrated in Fig. 5) decodes a spatio-temporal tube as the final prediction.

former encoder at both the frame level and a global level to perform better cross-modal alignment, which is followed by a template-based module designed for alleviating the prediction inconsistency problem faced with previous work. Despite the impressive progress, existing work still struggles to balance spatial and temporal cues for making more precise predictions. In this paper, we perform comprehensive space-time entanglement in the video encoding, the multi-modal fusion, and the tube decoding stages, which leads to fuller space-time interaction and hence better performance in terms of both spatial and temporal metrics.

**Vision-language understanding**. Inspired by the huge success of Transformers (Vaswani et al. 2017) in the natural language processing domain, many lines of research (Li et al. 2019; Su et al. 2019; Sun et al. 2019; Li et al. 2020; Kamath et al. 2021; Yang et al. 2021; Radford et al. 2021; Xu et al. 2021; Kim, Son, and Kim 2021; Yang et al. 2022b, 2023) extend the Transformer-based architecture and the pre-training paradigm (Devlin et al. 2019) to visual-linguistic tasks. Some (Li et al. 2020; Kim, Son, and Kim 2021; Li et al. 2023a; Liu et al. 2023) propose to pre-train Transformer-based models on large-scale image-text datasets to align visual-linguistic information, leading to substantial improvements in various down-stream tasks. Others (Sun et al. 2019; Li et al. 2021; Xu et al. 2021; Li et al. 2023b; Ding et al. 2022; Botach, Zheltonozhskii, and Baskin 2022; Hui et al. 2021) extend the Transformer framework to tackle video-language understanding tasks. In this paper, we focus on simultaneously modeling spatial and temporal representations in the context of cross-modal understanding for videos and text in the STVG task.

## Method

The overall pipeline of our proposed model, CoSTA, is illustrated in Fig. 3, which consists of a space-time feature encoding stage and a space-time entangled decoding stage. Given an untrimmed video, $\mathcal{V}$, and a language query, $\mathcal{S}$, the aim of our model is to output an object tube, $\mathbf{B} = \{b_i\}_{i=t_s}^{t_e}$, where $b_i = (x_i, y_i, w_i, h_i)$ indicates the spatial coordinates of the object, and $t_s$ and $t_e$ refer to the temporal boundaries of the event. In this section, we first present space-time feature encoding, then describe space-time entangled decoding. Finally, we detail the loss functions used for optimization.

### Space-Time Feature Encoding

The space-time feature encoding involves video encoding and multi-modal fusion. Intuitively, by mining space-time correlations in the video and aligning multi-modal features, the model can gain a deeper understanding of the events or actions referred to by the text. To achieve this, we first propose a space-time collaborative encoder to extract video features, then leverage a Transformer encoder to align video and language representations with global multi-modal temporal dependencies in mind.

**Space-time collaborative video encoding**. Leveraging 3D-CNN is a straightforward approach to obtain temporal information in videos (Tran et al. 2015; Carreira and Zisserman 2017; Gao et al. 2017; Tran et al. 2019; Zhang et al. 2021). However, it also introduces the problem of spatial misalignment, where temporal convolution or pooling pollutes single-frame representations (Botach, Zheltonozhskii, and Baskin 2022; Hui et al. 2021). Alleviating this problem is critical to addressing the STVG task, which demands precise spatial representations to localize the object. To this end, we propose a simple yet effective space-time collaborative encoder to learn video representations.

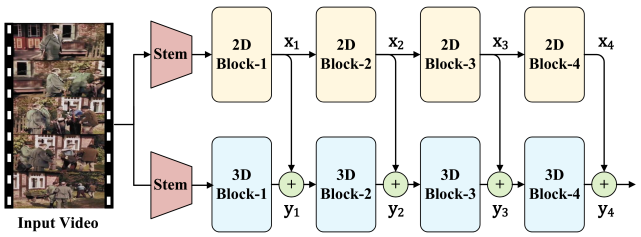


Figure 4: Feature map fusion in our space-time video encoder, where $x_l$ and $y_l$ are 2D feature maps and fused spatio-temporal feature maps from the $l$-th block, respectively.

Given an input video, $\mathcal{V} \in \mathbb{R}^{T\times H'\times W'\times C'}$, we first employ a 2D backbone (*e.g.*, ResNet) to extract four sets of multi-scale spatial feature maps $\{x_l\}_{l=1}^{4}$ for each frame, where $\boldsymbol{x}_l \in \mathbb{R}^{T\times H_l\times W_l\times C_l}$. Here, $l \in \{1,2,3,4\}$ indexes a stage in the backbone network. Then the spatial features are fused with temporal features modeled by a hierarchy of 3D CNN (*e.g.*, a 3D ResNet) layers. We denote the fused spatio-temporal feature maps as $\{y_l\}_{l=1}^{4}$, where $y_1$ is directly obtained by fusing first-stage outputs, and the rest of fusion is operated as follows:

$$y_l = \phi_l(y_{l-1}) \odot x_l, l \in \{2,3,4\}\,, \tag{1}$$

where $\phi_l$ denotes four stages of the 3D backbone, $x_l$ denotes spatial feature maps, '$\odot$' denotes the connection of two pathways. We keep the temporal stride as 1 so that the output feature maps are not temporally downsampled. As empirically detailed in the ablation studies, we have experimented with different connection functions and found that using the sum with feature normalization works the best for our model. We also detail the empirical study of alternative designs in the supplementary material.

Finally, we employ the last spatio-temporal feature maps $y_4$ as video features, which are further projected and flattened to yield the visual input of our multi-modal Transformer, denoted as $\mathcal{F}_v \in \mathbb{R}^{T\times HW\times C}$.

**Space-time multi-modal fusion**. After obtaining the video features, we employ a stack of Transformer encoder layers to fuse them with linguistic features, $\mathcal{F}_l \in \mathbb{R}^{T\times N\times C}$, which are obtained from a pre-trained language model (*e.g.*, RoBERTa (Liu et al. 2019c)) with an appended linear projection layer. The linguistic features are repeated $T$ times to align with the $T$ frames.

Inspired by existing work on video Transformers (Arnab et al. 2021; Jin et al. 2022), we perform factorized spatio-temporal encoding, which also corresponds to the "factorized encoder" variant in (Arnab et al. 2021). First, we concatenate the visual features with linguistic features at the sequence dimension. Then, we add $T$ learnable [CLS] tokens $\{c_i\}_{i=1}^{T}$ (one per frame) to obtain the input to the spatial self-attention layer, yielding features denoted as $\mathcal{F}_s \in \mathbb{R}^{T\times(HW+N+1)\times C}$. Notably, since the features are spatio-temporal, the self-attention layers perform motion-aware alignment between the video and text inputs. To model long-term temporal dependencies, we concatenate the output [CLS] tokens to form the input to the temporal interaction layers, denoted as $\mathcal{F}_t \in \mathbb{R}^{T\times C}$. The spatio-temporal multi-modal fusion is performed for $N$ times, encoding the video-language input into contextualized features, $\mathcal{F}_{out} \in \mathbb{R}^{T\times(HW+N)\times C}$, for decoding. Meanwhile, we also obtain a series of updated [CLS] tokens, $\{c_i\}_{i=1}^{T}$, which model global multi-modal context for individual frames and are further utilized for query generation in the next step.

## Space-time Entangled Decoding

In order to predict the target "object tube" from the encoded contextual features, previous work (Yang et al. 2022a; Jin et al. 2022; Su, Yu, and Xu 2021) performs parallel decoding, where the spatial location of the object and temporal boundaries of the event are only aware of each other in a post-processing stage. The target tube is synthesized by heuristically truncating the box sequence based on the predicted temporal boundaries, which potentially introduces invalid object-event combinations. To address this, we propose a space-time entangled decoder that bridges temporal boundary prediction and spatial localization via "entangled queries," which explicitly align the event and the object. Specifically, we first perform temporal decoding, then generate entangled queries to decode the spatial location. The process is detailed in the following.

**Temporal boundary prediction**. The temporal decoder is built with a stack of Transformer decoder layers. Each layer mainly includes a self-attention sub-layer, a cross-attention sub-layer, and a feed-forward sub-layer. The prediction is based on a query-to-feature measure. As illustrated in the left side of Fig. 5a, the temporal queries $\mathbf{Q} = \{q_i\}_{i=1}^{T}$ are formed by adding projected [CLS] tokens $\{c_i\}_{i=1}^{T}$ with 1D sinusoidal time embeddings (Vaswani et al. 2017), which can be formulated as:

$$q_i = \mathbf{W}(\theta_t(c_i) + \text{PE}(i)), i \in 1,2,3,...,T, \tag{2}$$

where $\theta_t$ denotes an MLP projection, PE denotes the 1D positional encoding, and $\mathbf{W}$ denotes a linear projection. In each decoding layer, the temporal queries first go through inter-frame self-attention, then, each query only attends to the encoded features of the corresponding frame in intra-frame cross-attention. Finally, we apply a prediction head on the output of the last layer to obtain the start and end probabilities for each frame, denoted as $p_t \in [0,1]^2$. The frames with maximum and valid (*i.e.*, $\hat{t_e} > \hat{t_s}$) start or end probabilities are chosen as the temporal boundaries $[\hat{t_s}, \hat{t_e}]$.

**Entangled query generation**. As detailed in Fig. 5a, we bridge the temporal decoder with the spatial decoder via "entangled queries," which consist of an event part and an object part. The event query refers to the time boundary predicted by the temporal decoder, and the object part is a set of positional queries (Meng et al. 2021; Liu et al. 2022; Liang et al. 2023) that perform spatial localization. Such design enforces the spatial decoder to ground objects according to the correlation between the duration of the event and the spatial appearance of the object in each frame. In our study, we interpret the object query as 4D anchor boxes (Liu et al. 2022). Our proposed method is a general strategy, and most of the positional query variants (Meng et al. 2021; Liu et al. 2022)

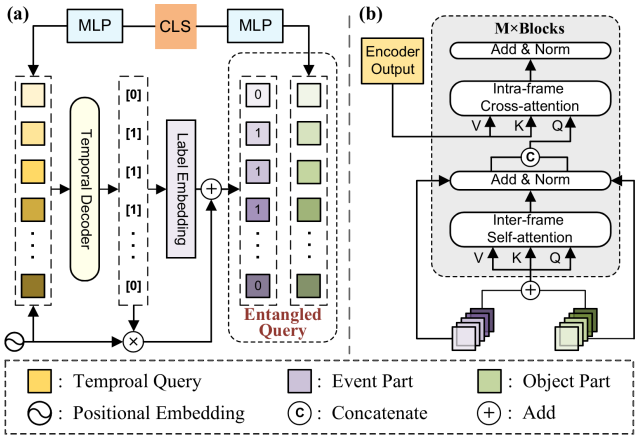


Figure 5: (a) Illustration of the spatio-temporal entangled decoder, and the entangled query generation process. The event part of the entangled query indicates whether the object appears in the current frame, while the object part contains the object's spatial information. (b) The entangled query updating mechanism in our spatial decoder.

of DETR-like decoders can be integrated to our method by being taken as the object part in our entangled query.

We start with the formulation of the event part. According to the time boundaries predicted previously, we harvest a set of binary labels $\{a_i\}_{i=1}^{T}$, which indicate whether each frame belongs to the target event and can be formulated as:

$$a_i = \begin{cases} 1, & \text{if } i \in [\hat{t_s}, \hat{t_e}], \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Then we embed each label via a learnable embedding layer. To model the duration of the event, 1D sinusoidal embedding is performed within time boundary $[\hat{t_s}, \hat{t_e}]$. The event part of the entangled query is finally formulated as follows:

$$\text{Tem}_i = \begin{cases} \text{Emb}(a_i) + \text{PE}(i), & \text{if } i \in [\hat{t_s}, \hat{t_e}], \\ \text{Emb}(a_i), & \text{otherwise.} \end{cases} \tag{4}$$

Due to the similar role between our object part and the positional query in DETR-like models (for spatial localization), following (Liu et al. 2022; Jin et al. 2022), we view them as 4D anchor boxes, which are initialized as:

$$\text{Anc}_i = \text{SE}(\sigma(\theta_s(c_i))), \tag{5}$$

where $\theta_s$ is MLP projection with 4 output channels, $\sigma(\cdot)$ denotes the sigmoid function, and SE denotes 2D sinusoidal positional encoding (Carion et al. 2020). The event part and object part are then sent to the spatial decoder for spatial localization.

**Spatial localization**. Fig. 5b depicts the details of the interaction mechanism of our entangled query in the spatial decoder. In each decoder layer, the event and object parts are first summed up for inter-frame self-attention, then concatenated for intra-frame cross-attention. The event query and object query aggregate the context features simultaneously, hence the anchors are updated with awareness of the relationship between the event and the target object. The anchor renewed by the last layer are the final spatial location of the objects, $b_t \in [0,1]^4$, within the time boundaries, $t \in [t_s, t_e]$. Finally, the generated bounding box sequence forms the output object tube, $\tilde{\mathbf{B}} \in [0,1]^{4(t_e - t_s + 1)}$.

**Training strategy**. At the training stage, we employ a sampling mechanism with ratio $\beta \in [0,1]$ to randomly select the time binary label $a_i$ from ground truth or from the temporal prediction, which leads to faster convergence.

## Loss Function

The model takes in a couple of videos and linguistic expressions, where each video is annotated with a set of bounding boxes, $\mathbf{B} \in [0,1]^{4(t_e - t_s + 1)}$, and the temporal boundaries, $[t_s, t_e]$. We use the sum of the Generalized IoU (Rezatofighi et al. 2019) loss and the L1 loss for spatial optimization:

$$\mathcal{L}_s = \lambda_0 \mathcal{L}_{\text{giou}}(B, \tilde{B}) + \lambda_1 \|B - \tilde{B}\|_1. \tag{6}$$

As for temporal supervision, following (Yang et al. 2022a; Jin et al. 2022; Rodriguez et al. 2020), we employ two 1D Gaussian distributions, $\pi_s$ and $\pi_t$, to represent "soft" supervision signals, and use the Kullback-Leibler divergence loss to measure the distance between the prediction and ground-truth distributions. The loss is formulated as follows:

$$\mathcal{L}_t = \mathcal{L}_{KLs}(\pi_s, \tilde{\pi}_s) + \mathcal{L}_{KLe}(\pi_e, \tilde{\pi}_e). \tag{7}$$

Lastly, we also supervise the "attendance" of the target object in each frame, which is predicted from the [CLS] tokens from the encoder and the queries updated by the last decoder layer, denoted as $\mathcal{L}_e$ and $\mathcal{L}_d$, respectively. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_s + \lambda_2 \mathcal{L}_t + \lambda_3 \mathcal{L}_e + \lambda_4 \mathcal{L}_d, \tag{8}$$

where the different $\lambda$s are balancing weights.

# Experiments

## Datasets and Metrics

**Datasets**. We evaluate our proposed method on two mainstream benchmarks **HC-STVG** (Tang et al. 2021) and **VidSTG** (Zhang et al. 2020c), which are annotated with both spatial bounding boxes and temporal boundaries. HC-STVG dataset consists of 5,660 videos in human-centric scenarios with annotated spatio-temporal tubes of the target person and is divided into training and test subsets with 4,500 and 1,160 video-sentence pairs. This dataset is extended to HC-STVG V2 with added data and cleaned labels, which contains 10,131 and 3,482 videos in training and validation subsets, respectively. VidSTG dataset totally consists of 99,943 sentence-tube pairs, where the language descriptions can be categorized into 44,808 declarative sentences and 55,135 interrogative sentences, with 79 types of various objects queried in 6,924 videos. Annotations in VidSTG are divided into training, validation and test subsets with 80,684, 8,956 and 10,303 distinct sentence-tube pairs corresponding to 5,436, 602 and 732 videos.

**Evaluation metrics**. We follow (Zhang et al. 2020c; Yang et al. 2022a; Jin et al. 2022) and choose m_vIoU and vIoU@R as the evaluation criteria. The vIoU metric is defined as $\text{vIoU} = \frac{1}{|S_u|} \sum_{t \in S_i} \text{IoU}\left(\hat{b}_t, b_t\right)$, where $S_u$ and $S_i$

| Methods | Params/M | HC-STVG V1 | | | |
|---|---|---|---|---|---|
| | | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
| STGVT | — | — | 18.15 | 26.81 | 9.48 |
| STVGBert | — | — | 20.42 | 29.37 | 11.31 |
| TubeDETR | 185.7 | 43.70 | 32.40 | 49.80 | 23.50 |
| STCAT | 207.8 | 49.44 | 35.09 | 57.67 | 30.09 |
| CoSTA (Ours) | 235.0 | **52.85** | **38.97** | **63.10** | **38.19** |

Table 1: Performance comparisons of the state-of-the-art on the HC-STVG v1 test set (%).

| ST-Encoder | ST-Decoder | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|---|
| ✓ | ✓ | **52.85** | **38.97** | **63.10** | **38.19** |
| ✓ | | 51.34 | 36.39 | 61.33 | 36.14 |
| | ✓ | 47.59 | 34.45 | 55.69 | 28.34 |

Table 2: Ablation results of spatio-temporal entanglement on the HC-STVG v1 test set.

| | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|
| **(a)** 2D/3D visual backbones | | | | |
| 2D backbone only | 50.09 | 37.57 | 60.42 | 34.41 |
| 3D backbone only | 52.59 | 36.08 | 59.10 | 32.67 |
| 2D & 3D backbones | **52.85** | **38.97** | **63.10** | **38.19** |
| **(b)** connection between video backbone | | | | |
| concatenate | 48.62 | 32.67 | 59.89 | 32.69 |
| Hadamard product | 49.12 | 34.39 | 60.93 | 34.36 |
| sum w. feat. norm. | **52.85** | **38.97** | **63.10** | **38.19** |
| **(c)** ratio of temporal ground truth for spatial decoding | | | | |
| 0% | 51.94 | 37.54 | 62.55 | 37.23 |
| 30% | 52.37 | 38.41 | 62.78 | 37.69 |
| 50% | 52.79 | 38.80 | 63.04 | 37.94 |
| 90% | 52.85 | **38.97** | **63.10** | **38.19** |
| 100% | **52.91** | 38.89 | 63.01 | 38.10 |

Table 3: Ablation studies on the HC-STVG v1 test set.

are the temporal union and intersection between the groung-truth tubes and the predicted ones. Besides, $b_t$ and $\hat{b}_t$ indicate the ground-truth bounding box and the predicted bounding box at frame $t$. The m_vIoU score is the average of vIoU scores over all samples in the test set. Moreover, vIoU@R is defined as the percentage of test samples whose vIoU>R among all videos in the test set. Additionally, we also adpot tIoU and sIoU to individually evaluate the temporal or spatial grounding accuracy of our model, where the former is defined as the temporal IoU measured between the ground-truth tubes and the predicted tubes $S_i/S_u$, while the latter one is calculated as the average of all IoUs between labeled bounding boxes and the predictions within ground-truth time boundary.

## Comparison with State-of-the-arts

In Tab. 1 and Tab. 4, we compare CoSTA with state-of-the-art methods on two mainstream benchmarks HC-STVG (Tang et al. 2021) and VidSTG (Zhang et al. 2020c) for STVG task. Notably, our method attains the best performance among all evaluation metrics across all datasets. Specifically, on HC-STVG V1 our model outperforms the second-best method STCAT (Jin et al. 2022) on m_vIoU and vIoU@0.5 metrics by large margins of 3.88%/8.10%. Meanwhile, our approach also shows its superiority over all existing methods on VidSTG across all performance metrics in Tab. 4. As the object queried in interrogative sentences, our CoSTA attains 2.17%/1.38% absolute improvement on m_vIoU and vIoU@0.5 metrics comparing to the second-best method (Jin et al. 2022). It is worth noting that our CoSTA exhibits greater improvement when dealing with HC-STVG dataset, which involves longer sentences with diverse events. This further demonstrates that our model excels better at entangling spatial and temporal information in complex multi-modal contexts.

## Ablation Study

In this section, we conduct extensive ablation experiments to first study the impact of performing various degrees of space-time entanglement in our framework, then ablate the alternative implementation of our CoSTA. The experiments are conducted on the test set of the HC-STVG v1 dataset.

**Impact of the spatio-temporal entanglement.** In this section, we ablate the impact of space-time entanglement at the feature encoding stage (ST-encoder) and the tube decoding stage (ST-decoder) in Tab. 2. We first ablate the impact of ST-encoder by replacing our video encoder with a 2D ResNet-101 backbone and removing the temporal encoding layers in our space-time multi-modal fusion Transformer. As shown in the third line of Tab. 2, this leads to a large drop of 5.26%, 7.41% and 9.85% on tIoU, vIoU@0.3 and vIoU@0.5 metrics, respectively. To ablate our ST-decoder, we remove the entangled query between temporal and spatial decoder, hence the spatial encoder is not aware of the time boundary of the event and the spatial query is directly initialized from [CLS] token. We can observe that the removal of the space-time entanglement in decoding stage leads to a drop of 1.77% and 2.05% in vIoU@0.3 and vIoU@0.5, demonstrating the advantages of our space-time entangled decoder.

**Components of the video encoder.** We empirically study the components of our visual encoder in Tab. 3a. It is worth noting that our full visual encoder incorporating both backbones outperforms each individual one on all evaluation metrics. We can observe that employing the 3D backbone leads to an improvement of 2.50% on m_tIoU compared to the 2D backbone. However, due to the spatial misalignment in single-frame representation, it underperforms the 2D-backbone in terms of m_vIoU, vIoU@0.3 and vIoU@0.5. Our visual encoder capitalizes on the strengths of both 2D and 3D backbones, resulting in the best performance on both spatial and temporal metrics.

**Connections between visual backbones.** In Tab. 3b, we explore different fusion methods between every block of two visual backbones in our video encoder, including 1) first concatenating the spatial and temporal features then using a convolutional layer to reduce channel dimension, 2) Hadamard product between two types of feature map, and 3) summing up with feature normalization. We observe that

| Methods | Parameters/M | Declarative Sentences | | | | Interrogative Sentences | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
| STGRN (Zhang et al. 2020c) | — | 48.47 | 19.75 | 25.77 | 14.60 | 46.98 | 18.32 | 21.10 | 12.83 |
| STGVT (Tang et al. 2021) | — | — | 21.62 | 29.80 | 18.94 | — | — | — | — |
| OMRN (Zhang et al. 2020b) | — | 50.73 | 23.11 | 32.61 | 16.42 | 49.19 | 20.63 | 28.35 | 14.11 |
| STVGBert (Su, Yu, and Xu 2021) | — | — | 23.97 | 30.91 | 18.39 | — | 22.51 | 25.97 | 15.95 |
| TubeDETR (Yang et al. 2022a) | 185.7 | 48.10 | 30.40 | 42.50 | 28.20 | 46.90 | 25.70 | 35.70 | 23.20 |
| STCAT (Jin et al. 2022) | 207.8 | 50.82 | 33.14 | 46.20 | 32.58 | 49.67 | 28.22 | 39.24 | 26.63 |
| CoSTA (Ours) | 235.0 | **52.08** | **35.09** | **48.44** | **34.03** | **51.84** | **29.86** | **41.31** | **28.02** |

Table 4: Performance comparisons of the state-of-the-art on the VidSTG test set (%).

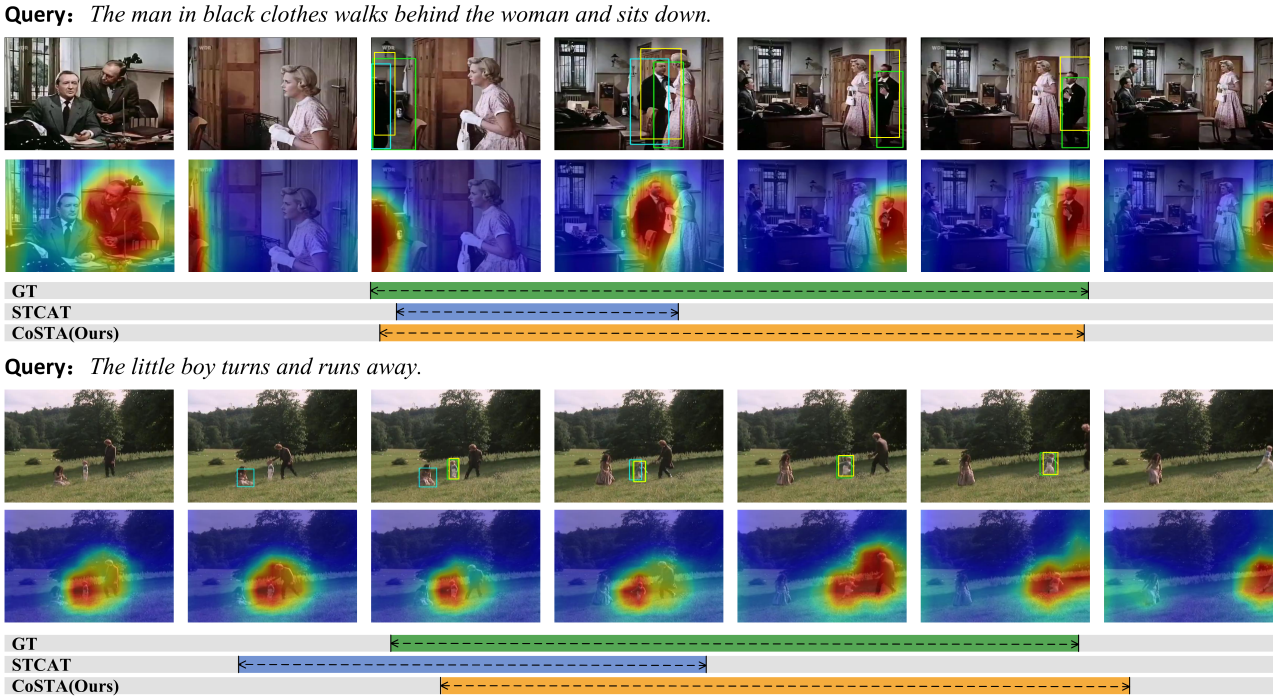


Figure 6: Visualized examples (Tang et al. 2021) of the spatio-temporal tubes and corresponding attention maps of the ground truth (green) and predictions produced by our approach (yellow) in comparison with STCAT (Jin et al. 2022) (cyan).

combining the spatial and temporal features by summing with normalization performs best in our space-time collaborative video encoder.

**Ratio of the temporal ground truth in training.** In Tab. 3c, we ablate the ratio of the ground truth that employed in the entangled query generation at training stage. Generally, the performance improves as more ground truth data is incorporated into the training process. We observe that the correct time boundary can help the model perform event-object alignment. Notably, since the impact of exposure bias (Zhang et al. 2019a), incorporating all ground truth leads to a decrease of 0.09% in vIoU@0.5 comparing to 90% ground truth, indicating that balancing the gap between training and inference is in demand for our entangled query.

## Visualization

In Fig. 6, we visualize some challenging scenarios and qualitatively compare the predicted tube of CoSTA with STCAT (Jin et al. 2022) on HC-STVG dataset. Our method gains more accurate time boundaries of the events in both scenarios. Furthermore, in the scene depicted at the bottom, we observe that STCAT suffers from inconsistency object prediction. It struggles with correctly aligning the "run away" event with the little boy presented in different frames while CoSTA identifies the same object over frames.

## Conclusion

In this paper, we have presented a comprehensive space-time entanglement framework for tackling the task of spatio-temporal video grounding, which addresses the issue of insufficient space-time interaction in existing methods. Extensive experiments have demonstrated its advantage with respect to state-of-the-art methods.

## Acknowledgements

This work of Yansong Tang is supported in part by National Natural Science Foundation of China (Grant No. 62206153), Shenzhen Key Laboratory of Ubiquitous Data Enabling (Grant No. ZDSYS20220527171406015) and Young Elite Scientists Sponsorship Program by CAST (No. 2024QNRC003). The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202 and the Shenzhen Science and Technology Program under Grant KQTD20170810150821146.

## References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*, 5803–5812.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. ViViT: A Video Vision Transformer. In *ICCV*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*.

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Chen, J.; Bao, W.; and Kong, Y. 2020. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In *ACMMM*.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020. Rethinking the bottom-up framework for query-based video localization. In *AAAI*, volume 34, 10551–10558.

Chen, Z.; Ma, L.; Luo, W.; and Wong, K.-Y. K. 2019. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *ICCV*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *CVPR*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*, 5267–5275.

Huang, D.-A.; Buch, S.; Dery, L.; Garg, A.; Fei-Fei, L.; and Niebles, J. C. 2018. Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In *CVPR*.

Hui, T.; Huang, S.; Liu, S.; Ding, Z.; Li, G.; Wang, W.; Han, J.; and Wang, F. 2021. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*, 4187–4196.

Jin, Y.; Yuan, Z.; Mu, Y.; et al. 2022. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *NIPS*, 35: 29192–29204.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 5583–5594. PMLR.

Li, G.; Duan, N.; Fang, Y.; Gong, M. U.-V.; and Jiang, D. U.-V. 2019. A Universal Encoder for Vision and Language by Cross-modal Pre-training. arXiv 2019. *arXiv preprint arXiv:1908.06066*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, L.; Lei, J.; Gan, Z.; Yu, L.; Chen, Y.-C.; Pillai, R.; Cheng, Y.; Zhou, L.; Wang, X. E.; Wang, W. Y.; et al. 2021. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*.

Li, M.; and Sigal, L. 2021. Referring Transformer: A One-step Approach to Multi-task Visual Grounding. In *NeurIPS*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*. Springer.

Liang, Y.; Yang, Z.; Tang, Y.; Fan, J.; Li, Z.; Wang, J.; Torr, P. H.; and Huang, S.-L. 2023. LUNA: Language as Continuing Anchors for Referring Expression Comprehension. In *ACMMM*.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019a. Learning to assemble neural module tree networks for visual grounding. In *ICCV*.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *ICLR*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019b. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019c.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional DETR for Fast Training Convergence. In *ICCV (ICCV)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*.

Rodriguez, C.; Marrese-Taylor, E.; Saleh, F. S.; Li, H.; and Gould, S. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2464–2473.

Shi, J.; Xu, J.; Gong, B.; and Xu, C. 2019. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *CVPR*.

Su, R.; Yu, Q.; and Xu, D. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 1533–1542.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. In *ICCV*.

Tan, R.; Plummer, B.; Saenko, K.; Jin, H.; and Russell, B. 2021. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *NIPS*.

Tang, Z.; Liao, Y.; Liu, S.; Li, G.; Jin, X.; Jiang, H.; Yu, Q.; and Xu, D. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8238–8249.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.

Tran, D.; Wang, H.; Torresani, L.; and Feiszli, M. 2019. Video classification with channel-separated convolutional networks. In *ICCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Yamaguchi, M.; Saito, K.; Ushiku, Y.; and Harada, T. 2017. Spatio-temporal person retrieval via natural language queries. In *ICCV*, 1453–1462.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022a. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 16442–16453.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving One-stage Visual Grounding by Recursive Sub-query Construction. In *ECCV*.

Yang, Z.; Tang, Y.; Bertinetto, L.; Zhao, H.; and Torr, P. H. 2021. Hierarchical interaction network for video object segmentation from referring expressions. In *BMVC*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022b. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2023. Semantics-Aware Dynamic Localization and Refinement for Referring Image Segmentation. *arXiv preprint arXiv:2303.06345*.

Zhang, H.; Sun, A.; Jing, W.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, 12870–12877.

Zhang, W.; Feng, Y.; Meng, F.; You, D.; and Liu, Q. 2019a. Bridging the gap between training and inference for neural machine translation. *arXiv preprint arXiv:1906.02448*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 655–664.

Zhang, Z.; Zhao, Z.; Lin, Z.; Huai, B.; and Yuan, N. J. 2020b. Object-aware multi-branch relation networks for spatio-temporal video grounding. *arXiv preprint arXiv:2008.06941*.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020c. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 10668–10677.