TubeRMC: Tube-conditioned Reconstruction with Mutual Constraints for Weakly-supervised Spatio-Temporal Video Grounding

Jinxuan Li[1], Yi Zhang[1], Jian-Fang Hu[1,2,3*], Chaolei Tan[4], Tianming Liang[1], Beihao Xia[5]

[1]Sun Yat-sen University,
[2] Guangdong Province Key Laboratory of Information Security Technology, China,
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China,
[4]The Hong Kong University of Science and Technology,
[5]Huazhong University of Science and Technology,
{lijx267,zhangy2799,liangtm9}@mail2.sysu.edu.cn, hujf5@mail.sysu.edu.cn, ctanak@connect.ust.hk, xbh_hust@hust.edu.cn

Abstract

Spatio-Temporal Video Grounding (STVG) aims to localize a spatio-temporal tube that corresponds to a given language query in an untrimmed video. This is a challenging task since it involves complex vision-language understanding and spatiotemporal reasoning. To eliminate reliance on fine-grained annotations like bounding boxes or temporal stamps, recent works have explored weakly-supervised setting in STVG. However, they typically follow a simple late-fusion manner, which generates tubes independent of the text description, often resulting in failed target identification and inconsistent target tracking. To address this limitation, we propose a Tube-conditioned Reconstruction with Mutual Constraints (**TubeRMC**) framework that generates text-conditioned candidate tubes with pre-trained visual grounding models and further refine them via tube-conditioned reconstruction with spatio-temporal constraints. Specifically, we design three reconstruction strategies from temporal, spatial, and spatio-temporal perspectives to comprehensively capture rich tube-text correspondences. Each strategy is equipped with a Tube-conditioned Reconstructor, utilizing spatio-temporal tubes as condition to reconstruct the key clues in the query. We further introduce mutual constraints between spatial and temporal proposals to enhance their quality for reconstruction. TubeRMC outperforms existing methods on two public benchmarks VidSTG and HCSTVG. Further visualization shows that TubeRMC effectively mitigates both target identification errors and inconsistent tracking.

Extended version — https://arxiv.org/abs/2511.10241

Introduction

STVG aims to predict a spatio-temporal tube (i.e., a sequence of bounding boxes within a specified time interval) corresponding to the event described by a language query in an untrimmed video. While existing methods (Yang et al. 2022; Gu et al. 2024) have achieved remarkable progress, they rely heavily on expensive tube-text annotations for supervised learning. To address this limitation, recent methods

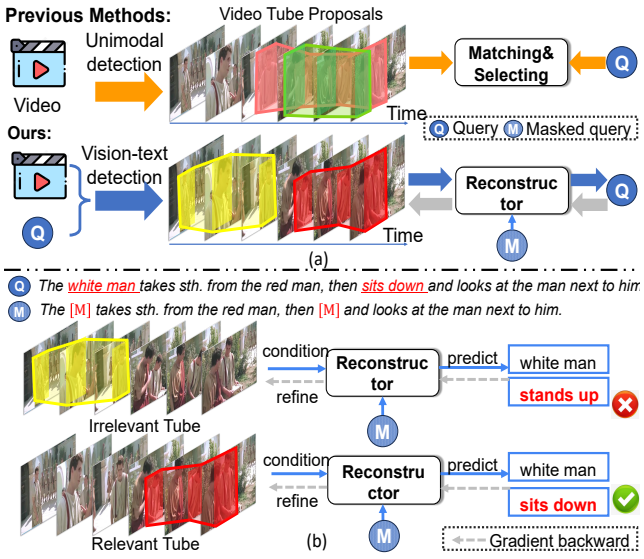

Figure 1: (a) Comparison with previous WSTVG methods. (b) Illustration of tube-conditioned reconstruction. The tube that matches the query descriptions can correctly reconstruct masked phrases. Moreover, the process can refine tubes predicted in detection. `[M]` means the masked phrase.

(Li et al. 2023; Jin and Mu 2024) have explored weakly-supervised STVG (WSTVG), which relies solely on video-text pairs without requiring bounding boxes or temporal annotations during training. These works typically follow a simple late-fusion manner, which employs a unimodal detector like Faster-RCNN (Ren et al. 2015) to produce tube proposals and then matches them with the input query, as shown in Figure 1(a). A main limitation of them is the tubes are generated independent of the text description, often resulting in failed target identification.(Li et al. 2025)

To overcome the limitation, we propose to introduce pre-trained visual grounding models, which can capture text-conditioned object localizations and provide more reliable spatial grounding results. A straightforward way to apply these models is concatenating frame-wise results to

*Corresponding author

form spatio-temporal tubes. However, this is ineffective for WSTVG since the object identification could be inconsistent across frames and temporal boundaries often fail to capture target event due to the lack of spatiotemporal understanding.

To fully unleash the potential of visual grounding models in WSTVG, we propose a novel framework **TubeRMC** that employs *tube-conditioned reconstruction* to enhance tube-text correspondence learning and event understanding, while utilizing *spatio-temporal mutual constraints* to refine both bounding boxes and temporal boundaries. The tube-conditioned reconstruction explores rich semantic dependencies between visual cues and textual descriptions by reconstructing masked text conditional on tube, learning the tube-text correspondences and achieving event understanding. In this way, the model learns to capture the tube-text correspondences by selecting the best tube proposal to reconstruct the text. As shown in Figure 1(b), the tube corresponding to the target event can reconstruct the key phrases 'white man' and 'sits down' in the masked sentence.

More specifically, instead of employing only temporal reconstruction like (Cao et al. 2023; Kim et al. 2024), we propose three reconstruction strategies from temporal, spatial, and spatio-temporal perspectives to capture more comprehensive tube-text correspondences. This is achieved by reconstructing key phrases from the text, conditioned on tube features generated under the guidance of 1D, 2D, and 3D Gaussian attention maps. Each strategy employs a Tube-conditioned Reconstructor, which leverages tube features to reconstruct the masked phrases from visual representations.

To further enhance the quality of tube proposals, we introduce mutual constraints consisting of: (1) a time-to-space constraint that ensures motion continuity for objects within the same scene, and (2) a space-to-time constraint that enforces temporal boundaries to align with high-confidence frames. These complementary constraints jointly improve object consistency and temporal prediction accuracy.

Extensive experiments on four public STVG benchmarks show significant improvement of TubeRMC over state-of-the-art (SOTA) methods. For example, in the HCSTVG-v1 dataset, our TubeRMC surpasses the previous SOTA method, VCMA (Jin and Mu 2024), by 4.74% in the m_vIoU metric. Moreover, compared to the baseline that directly using MDETR, TubeRMC demonstrates significant performance advantages. For instance, on the HCSTVG-v2 dataset, TubeRMC improves the baseline model by over 8% in terms of m_vIoU. These results highlight the effectiveness of our tube-conditioned reconstruction with mutual constraints framework. In summary, our contributions are:

- We propose a novel TubeRMC framework to learn fine-grained tube-text alignment without requiring annotations of bounding boxes or temporal intervals.
- We propose tube-conditioned reconstruction to comprehensively capture tube-text correspondences from 1D, 2D, 3D (temporal, spatial, spatio-temporal) perspectives, with mutual constraints to enhance proposal quality.
- Our approach outperforms previous state-of-the-art methods on VidSTG and HCSTVG benchmarks.

Related Work

Fully-supervised STVG. Spatio-Temporal Video Grounding focuses on identifying the target object both spatially and temporally based on a language query. Earlier methods (Zhang et al. 2020c, 2021; Tang et al. 2021) use detectors like Faster R-CNN to detect objects in each video frame and then ground the temporal locations based on the detection features. STGRN (Zhang et al. 2020c) and OMRN (Zhang et al. 2021) capture object relations through spatio-temporal graphs and multi-branch relation networks, respectively. Recent works (Yang et al. 2022; Tan et al. 2021a; Jin et al. 2022; Lin et al. 2023; Wasim et al. 2024; Liang et al. 2024; Wang et al. 2023; Gu et al. 2024; Liang et al. 2025b,a) have developed frameworks based on powerful pre-trained DETR-based (Carion et al. 2020) vision-language models with additional spatio-temporal interactions. These methods depend heavily on manual annotations, which is quite costly.

Weakly-supervised STVG. Many researchers propose to address Spatio-Temporal Video Grounding with less annotations (Chen et al. 2019; Chen, Bao, and Kong 2020; Li et al. 2023; Shi et al. 2019; Jin and Mu 2024) and develop weakly-supervised spatio-temporal grounding models. ADWS (Chen, Bao, and Kong 2020) provides a framework with mutually-guided spatio-temporal Multiple-Instance Learning to match each query with specific spatial regions in video frames. WINNER (Li et al. 2023) proposes a hierarchical video-language decomposition-alignment structure for multi-modal matching. VCMA (Jin and Mu 2024) uses the variational Expectation-Maximization algorithm to rebuild visual relationships between entities and achieve more accurate cross-modal alignment. However, they typically follow a detect-then-match process, where spatial proposals are generated by a unimodal detector like Faster-RCNN (Ren et al. 2015), tracked into tubes, and matched with text-based similarity. This ignores the unique correspondences between video regions and sentence components, which is crucial for WSTVG, limiting their performances. (Garg, Kumar, and Rawat 2025) directly employs G-DINO's (Liu et al. 2023) spatial detections to construct tubelets via tracking and applies progressive learning for temporal modeling. However, its performance is limited by inherited noise from G-DINO, as it cannot refine G-DINO's detection results.

Reconstruction for Weakly-supervised Video Grounding. Reconstruction-based methods (Lin et al. 2020; Zheng et al. 2022a; Cao et al. 2023; Kim et al. 2024; Zheng et al. 2022b) have been extensively studied for weakly-supervised video temporal grounding (WVTG) (Chen et al. 2022; Huang et al. 2021; Huang, Yang, and Sato 2023; Ma et al. 2020; Song et al. 2020; Tan et al. 2021b; Wang et al. 2021, 2022; Wang, Chen, and Jiang 2021; Zhang et al. 2020a,b). It assumes that temporal proposals matched well with the input text can reconstruct a sentence query from a randomly masked sentence query. (Zheng et al. 2022a) proposes a learnable Gaussian attention mask to generate proposals. It further introduces a mask-conditioned mechanism that restricts attention to the frames specified by the mask. (Kim et al. 2024) extends basic Gaussian proposals to Gaussian mixture models. However, these WVTG methods only focus on temporal masks and ignore the correspon-

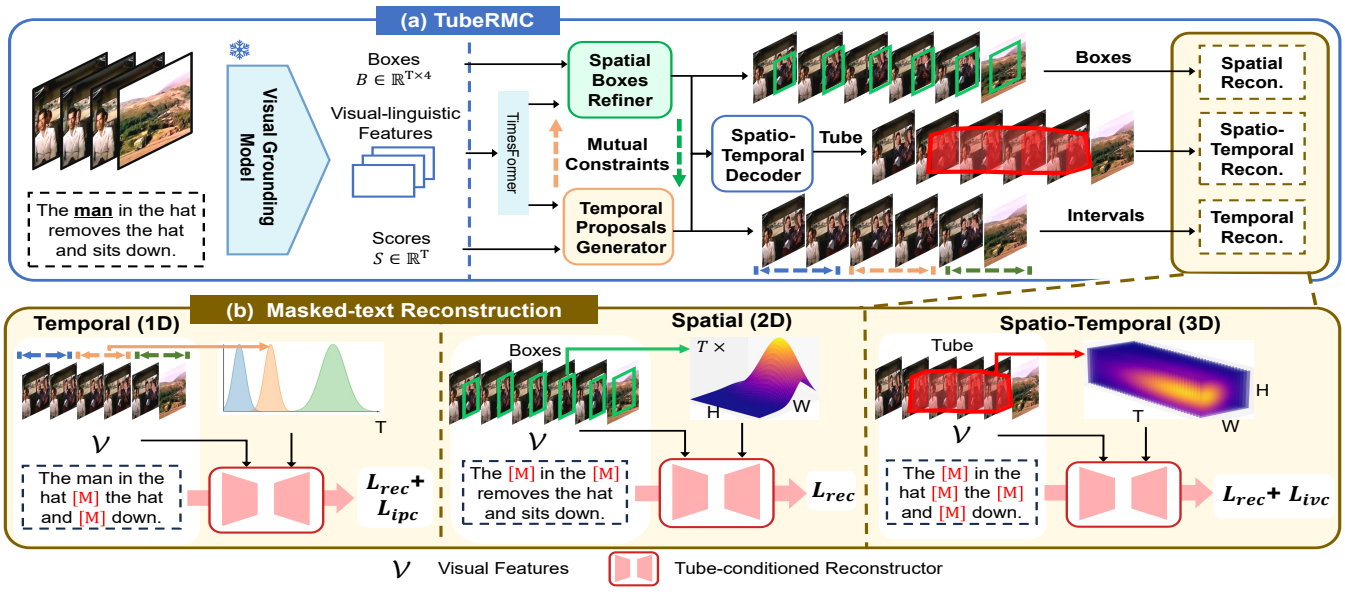

Figure 2: (a) TubeRMC Overview. It selects the most relevant bounding box for the subject token in each frame while extracting image-level visual-linguistic features (left), and generates temporal intervals (1D), spatial bounding boxes (2D), and spatio-temporal tube proposals (3D) for reconstruction (right). (b) Tube-conditioned Reconstruction Learning masks out phrases containing dynamic, static, or holistic event information in the sentence, then reconstructs the masked phrases condition on 1D, 2D, and 3D proposals. It further introduces mutual constraints between 1D and 2D proposals, employing time-to-space and space-to-time constraints to enhance spatio-temporal consistency in proposal generation.

dences between spatial information and text, which is crucial for STVG task. In this work, we propose a novel Tube-conditioned Reconstructor that can simultaneously handle both temporal and spatial attention masks.

Methodology

Model Overview

In Figure 2, we present our TubeRMC. We first introduce the **model architecture** of TubeRMC. It first employs a pre-trained visual grounding model to extract image-text correspondences and spatial grounding results per frame, and then performs cross-frame modeling to capture spatio-temporal context and generate diverse proposals for reconstruction. To learn fine-grained tube-text alignment without spatial or temporal annotations, we propose **Tube-conditioned Reconstruction Learning**. This learning scheme primarily incorporates three novel reconstruction strategies that rebuild masked phrases from temporal, spatial, and spatio-temporal perspectives respectively, comprehensively capturing spatio-temporal correspondences between visual cues and query descriptions. Furthermore, we introduce mutual constraints to enhance proposal quality for the reconstruction task.

Model Architecture

Static Cross-modal Extraction. We employ a pre-trained visual grounding model to extract frame-wise cross-modal representations and grounding results including bounding boxes and confidence scores. Recent visual grounding models MDETR(Kamath et al. 2021), G-DINO adopt the DETR-based (Carion et al. 2020) architecture, including an image backbone, a text encoder, and a transformer encoder-decoder with a box head and a confidence score head. These models take an image and a text query as input and output several bounding boxes with confidence scores for each query token. For each image, we rank all boxes by scores for the subject token and pick the highest-scoring box as the prediction for the whole query. We then concatenate the selected box in each frame and form a bounding box tube $B \in \mathbb{R}^{T\times 4}$ with confidence score vector $S \in \mathbb{R}^{T\times 1}$ where T is the number of frames. For target-invisable frames, MDETR tends to produce low-confidence regions, owing to its ability to align visual regions with text. These confidence scores can be utlized in our framework to filter out non-target regions.

Spatio-Temporal Modeling. The cross-modal features obtained from the grounding model are fed into the cross-modal TimesFormer (Bertasius, Wang, and Torresani 2021; Lin et al. 2023), obtaining the cross-modal features $F_t \in \mathbb{R}^{T\times(H\times W+L)\times d}$ and the global frame features $F_g \in \mathbb{R}^{T\times d}$. H and W denote the height and width of video frames, respectively, while L represents the length of input query. These features are then fed into Spatial Boxes Refiner and Temporal Proposals Generator to refine spatial grounding results and generate temporal proposals, respectively. In the Spatial Boxes Refiner, the object features (queries) of selected boxes from MDETR are processed with F_g via a cross-modal attention mechanism to capture inter-frame contextual relationships. The output queries are passed through a regression head to predict offsets that refine MDETR's outputs, yielding spatial proposal boxes and

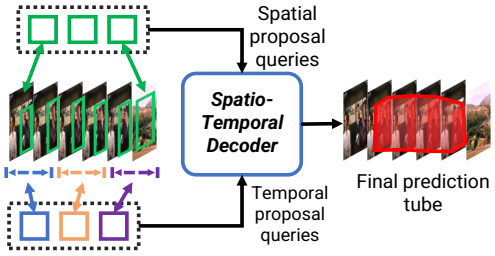

Figure 3: The effect of the Spatio-Temporal Decoder.

refined confidence scores. In Temporal Proposals Generator, we use K learnable queries in a temporal decoder to model temporal context. The architecture of decoder is similar to that of the spatial refiner, with key difference being that both F_t and F_g are used simultaneously in the cross-attention mechanism. The output embeddings are then passed through a fully-connected layer to predict temporal proposals.

The above pipeline provides basic spatial and temporal contextual modeling separately. However, since WSTVG task requires global video-text correspondences rather than individual spatial or temporal predictions, we further propose a Spatio-Temporal Decoder to integrate both spatial and temporal information and perform event-level predictions as shown in Figure 3. The spatial or temporal visual context in each proposal, embedded in spatial or temporal queries, is aggregated by a spatio-temporal query for global correspondence learning. We then use a spatio-temporal tube head to obtain final prediction. The spatial queries are employed for spatial localization, while the spatio-temporal query generates temporal boundaries.

Tube-conditioned Reconstruction Learning

Unlike previous reconstruction-based methods (Zheng et al. 2022a; Cao et al. 2023; Kim et al. 2024) which only focus on temporal grounding, we develop a Tube-conditioned Reconstructor (TR) that simultaneously take temporal and spatial masks as inputs, facilitating the model learning rich spatio-temporal visual-linguistic correspondences. Based on TR, we introduce three reconstruction strategies (Temporal, Spatial, and Spatio-Temporal) to capture tube-text correspondences from 1D, 2D, and 3D perspectives.

Representing Spatio-Temporal Masks as Gaussians. To enable gradient back-propagation in TR, we transform temporal intervals, spatial boxes, and spatio-temporal tubes into Gaussian distributions. For temporal intervals in Temporal Proposals Generator, we use a 1D Gaussian function to get temporal attention mask $M_t \in \mathbb{R}^T$. For spatial boxes in Spatial Boxes Refiner, we apply 2D Gaussian function for each frame, treating the center coordinates as the mean and the width as the standard deviation to produce $M_b \in \mathbb{R}^{HW}$. For the spatio-temporal tube, we employ 3D Gaussian function, treating the tube center coordinates as the distribution center and the tube's length, width, and height as the standard deviations, formulating masks of size $T \times HW$.

Tube-conditioned Reconstructor. As shown in Figure 4(a), TR consists of a Tube-conditioned Encoder and a Masked-text Reconstructor. We feed frame features and global features from TimesFormer into the encoder to model spatio-temporal dependencies under the guidance of spatial and temporal masks. Specifically, we define Mask-Attention as:

$$M\text{-}att(Q, K, V, M) = (\text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}}) \bigotimes M)V, \quad (1)$$

where M is the temporal or spatial Gaussian mask and $\bigotimes$ means multiplying by each row. As illustrated in Figure 4(b), the Tube-conditioned Encoder consists of a local branch (left) that focuses on local temporal and spatial context, and a global branch (right) that complements the understanding of global temporal information. Following this architecture, the encoder focuses on the visual regions corresponding to the masks, reducing responses to irrelevant areas. Finally, we apply spatial pooling to the visual features output from the final layer, resulting in spatio-temporal tube features $F_T \in \mathbb{R}^{T \times d}$. Different combinations of temporal and spatial masks enable the encoder to focus on different regions of the tube in video, producing different F_T.

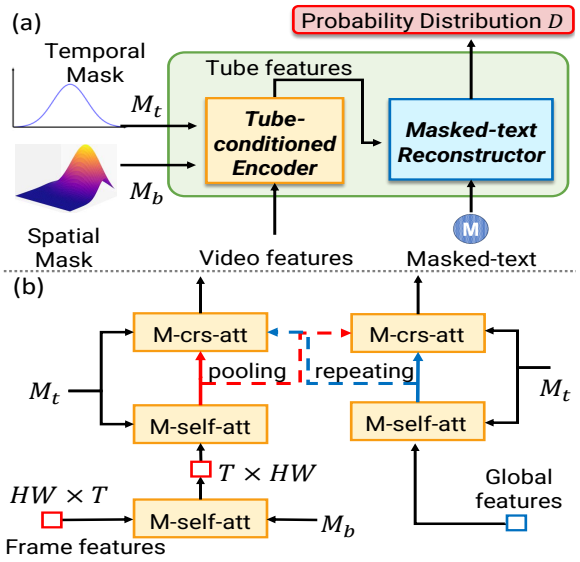

Figure 4: Architecture of Tube-conditioned Reconstructor: (a) Overview; (b) Tube-conditioned Encoder layer design.

The tube features F_T are then passed through the Masked-text Reconstructor. In the reconstructor, we first replace some word tokens in the original query with a special `[MASK]` token. The masked-text features are then fed into a transformer-decoder-like reconstructor to acquire the cross-modal semantic representations. Next, we apply a FFN with a softmax function to obtain the probability distribution D for the masked words in the vocabulary.

Reconstruction Strategies. To capture comprehensive spatio-temporal correspondences between visual cues and linguistic descriptions, we construct three types of reconstruction strategies including Temporal, Spatial, and Spatio-Temporal Reconstruction. For Temporal Reconstruction, predicates and their directly related nouns are selectively masked, which mainly includes motion-aware temporal in-

formation. The positive proposals are then transformed into 1D Gaussian temporal attention masks, respectively. For Spatial Reconstruction, we mask subject nouns with adjectives in the query and use the 2D spatial Gaussian Masks M_b to reconstruct the masked texts. This helps the model focus on the correspondence between object appearance information in both the video and the text. For Spatio-Temporal Reconstruction, we utilize 3D Gaussian masks for TR and randomly mask a set of words across sentence, with a higher probability of masking verbs, nouns, and adjectives. This enables the model to effectively capture spatio-temporally coupled correspondences.

Reconstruction Loss. We use cross-entropy loss $L_c(\cdot)$ (Lin et al. 2020; Kim et al. 2024) to measure the difference between the reconstructed query and the ground truth query. The reconstruction loss is formulated as:

$$L_{rec} = L_c(D_s) + L_c(D_t) + L_c(D_g), \tag{2}$$

where D_s, D_t, and D_g represent the predicted distribution from Spatial, Temporal and Spatio-Temporal Reconstruction, respectively.

Furthermore, to enable temporal visual-linguistic understanding without temporal annotations, we introduce a set of negative samples and design an inter-proposal contrastive loss for Temporal Reconstruction. For the construction of negative proposals, please refer to the next subsection. Since negative samples contain a significant amount of visual information unrelated to the query, their corresponding reconstruction loss should be higher than positive samples. Based on this, the inter-proposal contrastive loss L_{ipc} is defined as:

$$\begin{aligned} L_{ipc} = &\sum_{i=1, i\neq k^*}^{K} \max\big(L_c(D_t^{(k^*)}) - L_c(D_t^i) + \beta_1, 0\big) \\ &+ \sum_{i=1}^{K} \max\big(L_c(D_t^{(k^*)}) - L_c(D_{tn}^i) + \beta_2, 0\big), \end{aligned} \tag{3}$$

where k^* denotes the index of the positive proposal that minimizes the cross-entropy loss and D_{tn}^i means the predicted distribution for i-th negative temporal proposal.

Additionally, we generate two negative samples for Spatio-Temporal Reconstruction by perturbing the temporal prediction following (Zheng et al. 2022a). The easy negative sample uses an inverted temporal mask of the original tube, while the hard negative sample assigns uniform mask (all values set to 1) across the temporal dimension. The intra-video contrastive loss can be expressed as:

$$\begin{aligned} L_{ivc} = &\max\big(L_c(D_g) - L_c(D_{ghn}) + \beta_3, 0\big) + \\ &\max\big(L_c(D_g) - L_c(D_{gen}) + \beta_4, 0\big), \end{aligned} \tag{4}$$

where D_{ghn}, D_{gen} represent the reconstruction distributions of the hard and easy negative samples in Spatio-Temporal Reconstruction, respectively. The hyper-parameters β_1 to β_4 control the margins and $\beta_2 > \beta_1$, $\beta_4 > \beta_3$.

Mutual Constraints Learning. To improve both object consistency and temporal predictions quality for reconstruction, we propose mutual constraints learning. The space-to-Time constraint leverages spatial confidence scores to guide the generation of more precise temporal proposals. We first use scores $\hat{S}$ from Spatial Boxes Refiner to generate positive proposals. Specifically, the initial proposals are generated by selecting the top K scoring frames as temporal midpoints with a predefined width w_i. The Proposals Generator then predicts midpoint and width offsets, which are applied to yield the final positive proposals $\in \mathbb{R}^{K\times 2}$. Similarly, we construct negative temporal proposals using frames with the K-lowest scores. On this basis, the space-to-time constraint loss is defined to minimize overlaps both between different positive proposals and between positive and negative proposals. Furthermore, we propose the time-to-space constraint to make sure that objects maintain spatial continuity between adjacent frames within each scene. This is achieved by applying a loss that penalizes predicted boxes in adjacent frames with an IoU below a threshold within each temporal proposal.

Model Training and Inference

We train our model for weakly-supervised Spatio-Temporal Video Grounding in an end-to-end manner by minimizing the following loss function:

$$L_{total} = L_{rec} + L_{ipc} + L_{ivc} + L_{mc}, \tag{5}$$

where L_{mc} denotes mutual constraints loss, introduced to guide the refinement of both spatial and temporal proposals. During inference, we use the spatio-temporal tube output by Spatio-Temporal Decoder as the final grounding results.

Experiments

Experimental Settings

Datasets. We perform comprehensive experiments on the HCSTVG (Tang et al. 2021) and VidSTG (Zhang et al. 2020c) datasets. The HCSTVG-v1 consists of 4,500 video-text pairs for training and 1,160 for testing. The HCSTVG-v2 expands on v1 with improved annotation, containing 10,131 pairs for training, 2,000 for validation, and 4,413 for testing. Since the test annotations for HCSTVG-v2 are not publicly available, we report our results on the validation set. The VidSTG dataset includes 99,943 video-text pairs, comprising 44,808 declarative sentences and 55,135 interrogative sentences. The training, validation, and test sets consist of 80,684, 8,956, and 10,303 sentences respectively, along with 5,436, 602, and 732 videos respectively.

Evaluation Metrics. We follow (Zhang et al. 2020c; Li et al. 2023) and use mean vIoU as our main metric, which is calculated as: $vIoU = \frac{1}{|T_u|}\sum_{t\in T_i} IoU(\hat{b}_t, b_t)$, where T_i represents the intersection and T_u represents the union of the time intervals derived from the annotations and the predictions. $\hat{b}_t$ and b_t refer to the predicted and ground truth bounding boxes for the t-th frame, respectively. The vIoU score is averaged across all samples to obtain the mean vIoU (m_vIoU). Additionally, we report vIoU@R, which indicates the percentage of samples with a vIoU score greater than R. We also follow (Yang et al. 2022) and report the sIoU and tIoU metrics evaluating the spatial and temporal grounding performances in ablation studies, respectively. They are defined

Methods	HCSTVG-v1			VidSTG Declarative			VidSTG Interrogative		
	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
Two stages pipelines									
MATN+LCNet	4.41	3.53	1.12	8.16	8.03	3.59	6.97	6.64	3.05
GroundeR+CPL	5.23	4.18	1.25	8.28	8.35	3.68	7.16	7.28	3.23
One stage pipelines									
WSSTG	6.52	4.54	1.27	8.85	8.52	3.87	7.12	6.87	2.96
ADWS	8.20	4.48	0.78	8.96	7.86	3.10	8.57	8.64	2.88
Vis-Ctx	9.76	6.81	1.03	9.34	7.32	3.34	8.69	7.18	2.91
WINNER	14.20	17.24	6.12	11.62	14.12	7.64	10.23	11.96	5.46
VCMA	14.64	18.60	5.75	14.45	18.57	8.76	13.25	16.74	**7.66**
MDETR-Zero	12.26	18.10	5.03	11.13	14.41	6.06	6.52	7.80	2.96
MDETR+CPL	15.27	17.93	3.61	13.02	18.89	3.45	8.64	7.94	3.49
TubeRMC (Ours)	**19.38**	**23.88**	**6.75**	**15.93**	**25.16**	**9.09**	**13.47**	**18.79**	7.64

Table 1: Performance comparisons of the state-of-the-art methods on the VidSTG and the HCSTVG-v1 test set (%).

as:$sIoU = \frac{1}{|T_{gt}|}\sum_{t\in T_i} IoU(\hat{b_t}, b_t), tIoU = \frac{|T_i|}{|T_u|}$, where T_{gt} indicates the ground truth frame set.

Implementation Details. We use pre-trained MDETR with ResNet-101 (He et al. 2016) as the image encoder and Roberta-base (Liu et al. 2019) as the text encoder. K is set to 4 for HCSTVG and VidSTG. We set the number of transformer layers to 3 in Spatio-Temporal modeling, while TR utilizes a 6-layer architecture. The margin weights $\beta_1, \beta_2, \beta_3, \beta_4$ are set to 0.5, 0.7, 0.5, 0.7, respectively.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
MDETR-Zero	12.21	17.15	5.80
MDETR+CPL	15.09	24.95	6.50
TubeRMC (Ours)	**20.64**	**26.05**	**8.15**

Table 2: Performances on HCSTVG-v2 validation set (%).

Experimental Results

To verify the effectiveness of TubeRMC, we develop two baseline models based on MDETR and compare the performance of TubeRMC with the baselines and previous WSTVG methods on HCSTVG and VidSTG.

Comparison with previous SOTA. As shown in Table 1, TubeRMC surpasses previous SOTA two-stage and one-stage methods by a margin of 12.50% and 4.74% in m_vIoU in HCSTVG-v1, respectively. Furthermore, we also outperform all methods on VidSTG Declarative and Interrogative. These results strongly validate effectiveness of TubeRMC.

Comparison with baselines. MDETR-Zero is a zero-shot baseline that relies solely on MDETR predictions and generates temporal boundaries based on frame-wise confidence scores. It achieves performance comparable to the previous SOTA on HCSTVG-v1 and VidSTG Declarative without extra training. Moreover, we apply temporal reconstruction from WVTG to MDETR's predictions and propose a new baseline MDETR+CPL. As shown in Table 1, it surpasses VCMA (Jin and Mu 2024) by 0.63% in terns of m_vIoU on HCSTVG-v1. Although these two baselines achieve certain performance gains, the results remain unsatisfactory since they fail to capture spatio-temporal correspondences. Compared to them, TubeRMC demonstrates a significant performance advantage. For example, Table 2 shows that TubeRMC outperforms MDETR-Zero and MDETR+CPL by 8.43% and 5.55% in m_vIoU on HCSTVG-v2, respectively. Furthermore, due to the significant gap between the pretraining corpus of MDETR and VidSTG Interrogative, the two baselines exhibit suboptimal performance on this dataset. However, TubeRMC effectively mitigates this gap and achieves superior results compared to previous approaches. This demonstrates the effectiveness of our tube-conditioned reconstruction with mutual constraints design.

Foundation Model	Backbone	m_vIoU	vIoU@0.3	vIoU@0.5
Faster-RCNN	ResNet-101	15.31	19.97	6.06
MDETR	ResNet-101	19.38	23.88	6.75
G-DINO	Swin-T	19.47	24.12	7.03
G-DINO	Swin-B	21.15	26.81	8.11

Table 3: Evaluation on varied viusal grounding models.

Ablation Study

In this section, we conduct ablation experiments on HCSTVG-v1 dataset (Tang et al. 2021) to evaluate the effectiveness of some key components in the proposed method.

Ablation on varied visual grounding models. To study the effect of foundation models, we replace the visual grounding model MDETR with other models like G-DINO (Liu et al. 2023) and ViLBERT (Faster-RCNN) (Lu et al. 2019). The results are presented in Table 3.We can observe that replacing MDETR with a stronger pre-trained foundation model (e.g., G-DINO-Swin-B) can achieve better performance, improving m_vIoU from 19.38% to 21.15%. To strike a balance between speed and performance, we use MDETR by default.

Impact of the reconstruction. Our approach intends to construct three different masked-text reconstruction strategies. Here, we evaluate the effect of our Tube-conditioned Reconstruction by removing the strategies. The experimental results are tabulated in Table 4. As can be observed, either the Spatial or Temporal Reconstruction performs pos-

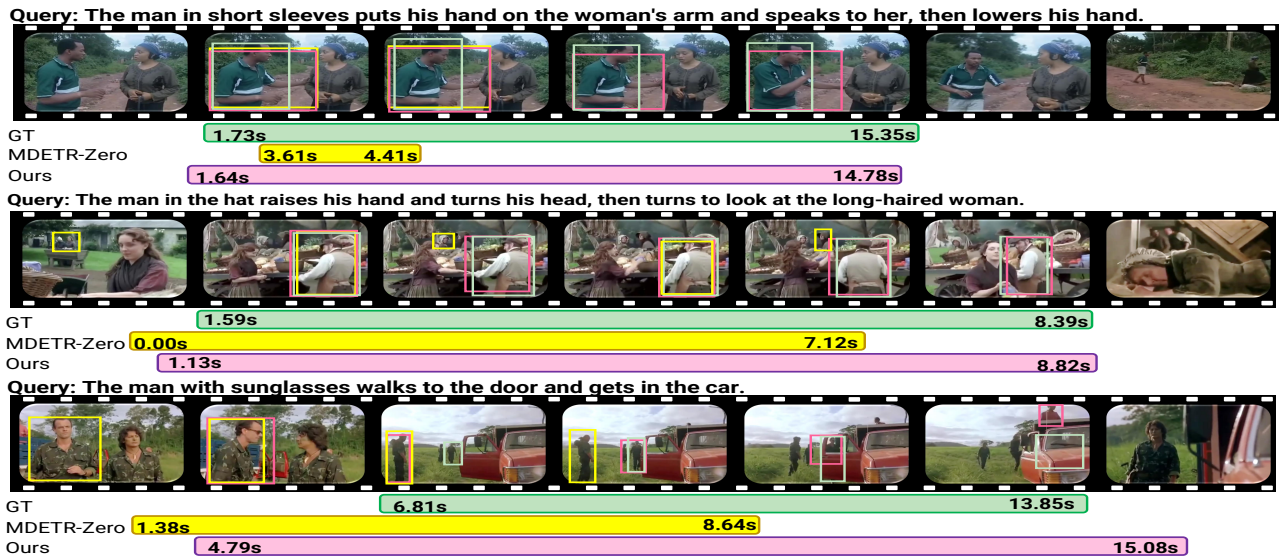

Figure 5: Visualization of spatio-temporal predictions for our method (pink), MDETR-Zero (yellow) and ground truth (green).

Spatial	Temporal	Spatio-Temporal	m_vIoU	vIoU@0.3	vIoU@0.5
×	×	×	14.91	14.87	2.32
✓	×	×	15.24	14.74	3.10
×	✓	×	17.59	20.25	4.74
✓	✓	×	18.12	20.43	5.00
×	×	✓	17.37	20.78	5.17
✓	✓	✓	**19.38**	**23.88**	**6.75**

Table 4: Evaluation on the reconstruction strategies.

ID	s-to-t	t-to-s	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_sIoU
(a)			15.87	11.98	1.90	26.69	59.83
(b)	✓		17.65	20.51	5.26	29.04	59.95
(c)		✓	17.07	14.06	2.32	27.38	60.13
(d)	✓	✓	**19.38**	**23.88**	**6.75**	**30.94**	**61.67**

Table 5: Evaluation on the mutual constraints.

itively to improve system performance. The performance is further improved when both the reconstructions are employed, which means that they can complement well with each other. Specifically, using the Temporal Reconstruction can improve the prediction accuracy by 2.68% in the term of m_vIoU, demonstrating the importance of temporal modeling in addressing WSTVG. As expected, combining all the strategies can obtain the best results, enabling the model to capture tube-text correspondences comprehensively.

Impact of mutual constraints. We investigate the influences of mutual constraints in Table 5. For experiment (a) and (b), the time-to-space constraint is not included (termed 't-to-s'). In experiments corresponding to row (a) and (c), we exclude the space-to-time constraint (termed 's-to-t'). As shown, our space-to-time constraint boosts the performance of m_tIoU (from 26.69 to 29.04 and from 27.38 to 30.94) and time-to-space constraint is beneficial for improving spatial grounding results. As expected, using both constraints further improves spatio-temporal prediction performance.

Visualization Results

We provide several visualization samples predicted by MDETR-Zero baseline and TubeRMC. As shown in Figure 5, MDETR provides unreliable temporal estimates (line 1) and unstable spatial predictions (line 2). In contrast, TubeRMC captures tube-text correspondences and provides accurate spatio-temporal predictions. This highlights the advantages of tube-conditioned reconstruction in WSTVG learning. Furthermore, we present a challenging case in line 3. The appearance of target 'man with sunglasses' is highly ambiguous during GT period. MDETR mistakenly assigns the bounding boxes to another man performing a similar action due to severe viewpoint changes and object occlusions. Although disrupted by MDETR, our TubeRMC still provides relatively accurate spatio-temporal predictions. We will keep overcoming this issue in our future work. A potential solution is to introduce additional tracking algorithms to guide the model in generating higher-quality tubes.

Conclusion

In this work, we propose a novel TubeRMC framework for WSTVG. TubeRMC first deploys a visual grounding model to extract image-text correspondences and then performs cross-frame modeling to capture spatio- temporal context. We propose Tube-conditioned Reconstruction Learning, incorporating three novel reconstruction strategies to capture tube-text correspondences from 1D, 2D, and 3D perspectives. Furthermore, we introduce mutual constraints to enhance proposal quality. Our approach achieves SOTA on VidSTG and HCSTVG, demonstrating the effectiveness of tube-conditioned reconstruction with mutual constraints.

Acknowledgements

This work was supported partially by the NSFC (U22A2095, 6247072922), Guangdong Natural Science Funds Project (2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B15-15020009), and Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026).

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Cao, M.; Wei, F.; Xu, C.; Geng, X.; Chen, L.; Zhang, C.; Zou, Y.; Shen, T.; and Jiang, D. 2023. Iterative Proposal Refinement for Weakly-Supervised Video Grounding. In *CVPR*, 6524–6534.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, J.; Bao, W.; and Kong, Y. 2020. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3789–3797.

Chen, J.; Luo, W.; Zhang, W.; and Ma, L. 2022. Explore inter-contrast between videos via composition for weakly supervised temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 267–275.

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*.

Garg, A.; Kumar, A.; and Rawat, Y. S. 2025. Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3384–3394.

Gu, X.; Fan, H.; Huang, Y.; Luo, T.; and Zhang, L. 2024. Context-Guided Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18330–18339.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; Liu, Y.; Gong, S.; and Jin, H. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7199–7208.

Huang, Y.; Yang, L.; and Sato, Y. 2023. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18908–18918.

Jin, Y.; Li, Y.; Yuan, Z.; and Mu, Y. 2022. Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding. *arXiv preprint arXiv:2209.13306*.

Jin, Y.; and Mu, Y. 2024. Weakly-Supervised Spatio-Temporal Video Grounding with Variational Cross-Modal Alignment. In *European Conference on Computer Vision*, 412–429. Springer.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kim, S.; Cho, J.; Yu, J.; Yoo, Y.; and Choi, J. Y. 2024. Gaussian Mixture Proposals with Pull-Push Learning Scheme to Capture Diverse Events for Weakly Supervised Temporal Video Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2795–2803.

Li, J.; Tan, C.; Chen, H.; Ma, J.; Hu, J.-F.; Zheng, W.-S.; and Lai, J. 2025. Image-to-Video Transfer Learning based on Image-Language Foundation Models: A Comprehensive Survey. *arXiv preprint arXiv:2510.10671*.

Li, M.; Wang, H.; Zhang, W.; Miao, J.; Zhao, Z.; Zhang, S.; Ji, W.; and Wu, F. 2023. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23090–23099.

Liang, T.; Jiang, H.; Yang, Y.; Tan, C.; Li, S.; Zheng, W.-S.; and Hu, J.-F. 2025a. Long-RVOS: A Comprehensive Benchmark for Long-term Referring Video Object Segmentation. *arXiv preprint arXiv:2505.12702*.

Liang, T.; Lin, K.-Y.; Tan, C.; Zhang, J.; Zheng, W.-S.; and Hu, J.-F. 2025b. Referdino: Referring video object segmentation with visual grounding foundations. *arXiv preprint arXiv:2501.14607*.

Liang, Y.; Liang, X.; Tang, Y.; Yang, Z.; Li, Z.; Wang, J.; Ding, W.; and Huang, S.-L. 2024. CoSTA: End-to-End Comprehensive Space-Time Entanglement for Spatio-Temporal Video Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3324–3332.

Lin, Z.; Tan, C.; Hu, J.-F.; Jin, Z.; Ye, T.; and Zheng, W.-S. 2023. Collaborative Static and Dynamic Vision-Language Streams for Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23100–23109.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*, volume 34, 11539–11546.

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, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. 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. *Advances in neural information processing systems*, 32.

Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; and Yoo, C. D. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 156–171. Springer.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10444–10452.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*.

Tan, C.; Lin, Z.; Hu, J.-F.; Li, X.; and Zheng, W.-S. 2021a. Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. *arXiv preprint arXiv:2106.10634*.

Tan, R.; Xu, H.; Saenko, K.; and Plummer, B. A. 2021b. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2083–2092.

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*, 1–1.

Wang, W.; Liu, J.; Su, Y.; and Nie, W. 2023. Efficient Spatio-Temporal Video Grounding with Semantic-Guided Feature Decomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4867–4876.

Wang, Y.; Deng, J.; Zhou, W.; and Li, H. 2021. Weakly supervised temporal adjacent network for language grounding. *IEEE Transactions on Multimedia*, 24: 3276–3286.

Wang, Y.; Liu, M.; Wei, Y.; Cheng, Z.; Wang, Y.; and Nie, L. 2022. Siamese alignment network for weakly supervised video moment retrieval. *IEEE Transactions on Multimedia*, 25: 3921–3933.

Wang, Z.; Chen, J.; and Jiang, Y.-G. 2021. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1459–1468.

Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2024. VideoGrounding-DINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18909–18918.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Lin, Z.; Zhao, Z.; Zhu, J.; and He, X. 2020a. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4098–4106.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020b. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33: 18123–18134.

Zhang, Z.; Zhao, Z.; Lin, Z.; Huai, B.; and Yuan, J. 2021. Object-Aware Multi-Branch Relation Networks for Spatio-Temporal Video Grounding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 9780999241165.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, M.; Huang, Y.; Chen, Q.; and Liu, Y. 2022a. Weakly supervised video moment localization with contrastive negative sample mining. In *AAAI*, volume 1, 3.

Zheng, M.; Huang, Y.; Chen, Q.; Peng, Y.; and Liu, Y. 2022b. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15555–15564.