Learning Feature Semantic Matching for Spatio-Temporal Video Grounding

Tong Zhang, Hao Fang, Hao Zhang, Jialin Gao, Xiankai Lu, *Member IEEE*, Xiushan Nie, Yilong Yin

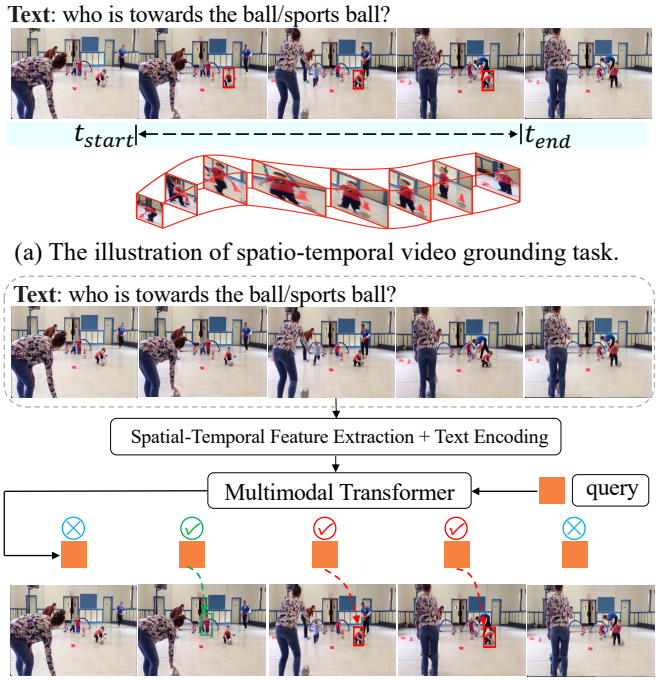
***Abstract*—Spatio-temporal video grounding (STVG) aims to localize a spatio-temporal tube, including temporal boundaries and object bounding boxes, that semantically corresponds to a given language description in an untrimmed video. The existing one-stage solutions in this task face two significant challenges, namely, vision-text semantic misalignment and spatial mislocalization, which limit their performance in grounding. These two limitations are mainly caused by neglect of fine-grained alignment in cross-modality fusion and the reliance on a text-agnostic query in sequentially spatial localization. To address these issues, we propose an effective model with a newly designed Feature Semantic Matching (FSM) module based on a Transformer architecture to address the above issues. Our method introduces a cross-modal feature matching module to achieve multi-granularity alignment between video and text while preventing the weakening of important features during the feature fusion stage. Additionally, we design a query-modulated matching module to facilitate text-relevant tube construction by multiple query generation and tubulet sequence matching. To ensure the quality of tube construction, we employ a novel mismatching rectify contrastive loss to rectify the mismatching between the learnable query and the objects corresponding to the text descriptions by restricting the generated spatial query. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods on two challenging STVG benchmarks. Code is publicly available at https://github.com/tongzhang111/acm-mm.**

***Index Terms*—Spatio-temporal video grounding, Multi-modal attention, Contrastive loss.**

I. Introduction

Video data has received more and more attention because it can better help the agent to understand scenes by using both the temporal and spatial context. Among them, video understanding has attracted many researchers in the past few decades, and has great application prospects in the fields of product retrieval and advertising recommendation [1], [2].

Compared to images and text, video conveys richer semantic knowledge, as well as more diverse and complex activities, such as video hyperlinking that aims to enhance the accessibility of large video datasets [3]. Recently, video language grounding has become a prominent and fundamental task in the multi-modal video understanding field. It aims to predict the timestamps from the video content specified with a given natural text. This task serves as a bridge between video and language and has gained significant attention in recent years [4]–[10]. Previous studies have mainly focused on extracting the temporal moments from videos, understanding the association of language with the temporal boundaries in videos is particularly important. However, understanding how language is associated with the spatial boundaries of text-relevant objects in videos has not been thoroughly explored.



(a) The illustration of spatio-temporal video grounding task.

(b) Inconsistent prediction from previous end-to-end method.

Fig. 1. (a) Illustration of spatio-temporal video grounding task. (b) Previous methods use one query to predict the bounding boxes based on DETR. Since a query is difficult to represent all objects and the disorder of predicted bounding boxes, it will cause the prediction results to be inconsistent with the target (Blue represents bounding boxes with empty predictions, green represents inconsistent prediction, and red represents the correct prediction.).

Therefore, a compound task, termed spatio-temporal video grounding (STVG) was introduced by [11] recently. As depicted in Fig. 1 (a), given an untrimmed video, STVG is to predict a sequence of bounding boxes (*i.e.*, spatio-temporal tube) that are relevant to the given query. Compared to temporal video grounding [12]–[14], STVG is more challenging as it requires the simultaneous localization of the temporal range of the activity described in the language and the corresponding object in spatial domain.

The majority of prevailing approaches [11], [15]–[18]

Tong Zhang, Hao Fang, Xiankai Lu, Yilong Yin, are with the School of Software, Shandong University, Jinan 250101,China.(E-mail: tz21@mail.sdu.edu.cn; carrierlxk@gmail.com; ylyin@sdu.edu.cn)

Hao Zhang is with the School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore.

Jialin Gao is with the AI Singapore, National University of Singapore, 117602, Singapore

Xiushan Nie is with Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd. and the School of Computer Science and Technology, Shandong Jianzhu University, Jinan China. (E-mail:niexsh@hotmail.com)

Corresponding author: Xiushan Nie, Yilong Yin.

streamline the spatio-temporal video grounding as a two-stage pipeline. They generate the object proposals or tubes via the pre-trained object detectors (*e.g.*, Faster RCNN [19]) in the first stage and then rank all the proposals or tubes by the similarity with text features in the second stage. Thus, the performance of these two-stage methods heavily relies on the quality of the pre-generated proposal. Video features extracted by the pre-trained object detector are usually intrinsic and have no correlation to the text description. Thus, the generated proposals are class-agnostic and contain lots of text-irrelevant proposals, which would degrade the proposal quality.

Inspired by the MDETR [20], several solutions [21], [22] seek the one-stage transformer encoder-decoder framework for spatio-temporal video grounding. Despite promising performance has been achieved, these approaches generally suffer from **cross-modal misalignment issue** and **incorrect predictions in spatial grounding**. Previous one-stage solutions [21], [23] neglect the text is inconsistent with the target. Therefore, previous cross-modal fusion strategies lead to important information loss and semantic misalignment. For example, given the query of "who holds the sports ball", the target "people" cannot be determined only from the text description. Therefore, only implementing cross-modal feature fusion may result in only enhancing the visual features corresponding to the sports ball and weakening the adult features. Although the STCAT [22] adopts the spatial and temporal interaction layer to obtain a better spatio-temporal feature representation, still does not solve the issue that only a part of the video corresponds to the text. Therefore, the STCAT also has the problem of misalignment of video and text features. The matching difficulty between object queries and corresponding targets is the major problem for DETR's architecture, which easily leads to mismatching. This issue is compounded in the cross-attention layer, as noted by Group-DETR [24]. A limited number of queries can expand the search range for each query, hindering the precise extraction of object features within specific regions. Moreover, for tube construction, existing methods (*e.g.*, TubeDETR [21]) typically use a learnable query in the DETR's architecture to predict a unique bounding box for each frame, regardless of whether or not the selected query corresponds to the language (see Fig. 1(b)). This approach can result in mismatching between the learnable query and target objects corresponding to the test descriptions. Therefore, the mismatching further leads to incorrect predictions and mislocalization.

To tackle the aforementioned issues, we develop the Feature Semantic Matching (FSM) . On top of the prevalent detection network TubeDETR [21], FSM introduces the cross-modal feature matching module to ensure video features and text features consistent and prevent some important spatial features weaken in the cross-modal fusion stage. this fusion module is transformer-based yet designed to emphasize correspondences between text and spatio-temporal information by multi-granularity alignment. At the frame-level, the fusion module incorporating the gate operations can align video and text features while highlighting the important spatial features. At the video-level, the module uses the similarity matrix to weigh the video features and uses a learnable scalar to further enhance the temporal features.

Moreover, to effectively associate the spatial query (bounding boxes) with the text description (*i.e.*, language), we propose a query-modulated matching module, which consists of transformer decoder, tubelet sequence matching, and mismatching rectify contrastive loss. Our model predicts multiple detection candidates instead of using a query and generating one box per frame (like TubeDETR [21] and STCAT [22]). Then, it employs tube sequence matching to find the optimal match between all queries and ground-truth. Considering the queries are temporally disordered across different frames, we further use the mismatching rectify contrastive loss to improve the association between the queries and the text description by contrastive learning. In this way, our model can tackle the difficult matching issue of previous methods and generate a more accurate spatial bounding box corresponding to the text.

The main contributions of this work are summarized as:

- We impose a cross-modal feature matching module to mitigate the video and text features inconsistency issue and preserve important features in the STVG task. In this way, the cross-modal features are enhanced to yield a more precise spatio-temporal localization.
- We firstly use multiple queries and propose query-modulated matching module to match a relative correct tube from multiple tubes. Then, we use the mismatching rectify contrastive loss module with few parameters to keep the object queries consistent with the target.
- We conduct comprehensive experiments on two challenging STVG benchmarks (*i.e.*, VidSTG [25] and HC-STVG [26]) further demonstrate that our method obtained new state-of-the-art performance.

II. Related Work

A. Video Grounding

The video grounding task can be classified into *temporal grounding* and the newly emerging *spatio-temporal grounding*. The former is a crucial multi-modal task aimed at localizing the timestamp of a video event based on a textual description. However, most temporal grounding works [7], [27]–[29] concentrate on designing architectures to extract temporal cues, overlooking the importance of spatial information.

In the context of temporal grounding tasks, most existing methods [13], [30] focus on performing fine-grained interaction between video and language. Meanwhile, the spatio-temporal video grounding aims at the intersection of spatial and temporal localization. Spatio-temporal video grounding demands information fusion between temporal-level features and text as well as the fusion between spatial-level features and text. Therefore, spatio-temporal video grounding is more difficult than temporal grounding. To obtain the location prediction in each frame, most existing STVG approaches [31], [32] rely on pre-extracted object proposals. Thus, the performance of these methods is heavily limited by the pre-trained object detector. If the object features extracted by the object detector have a huge domain gap with the text features, it is difficult to perform feature alignment. Recently, transformer-based methods have made great progress in object detection and multi-modal tasks. STVGBert [23] proposes a one-stage transformer-based approach that extends the VilBERT [33] to settle this

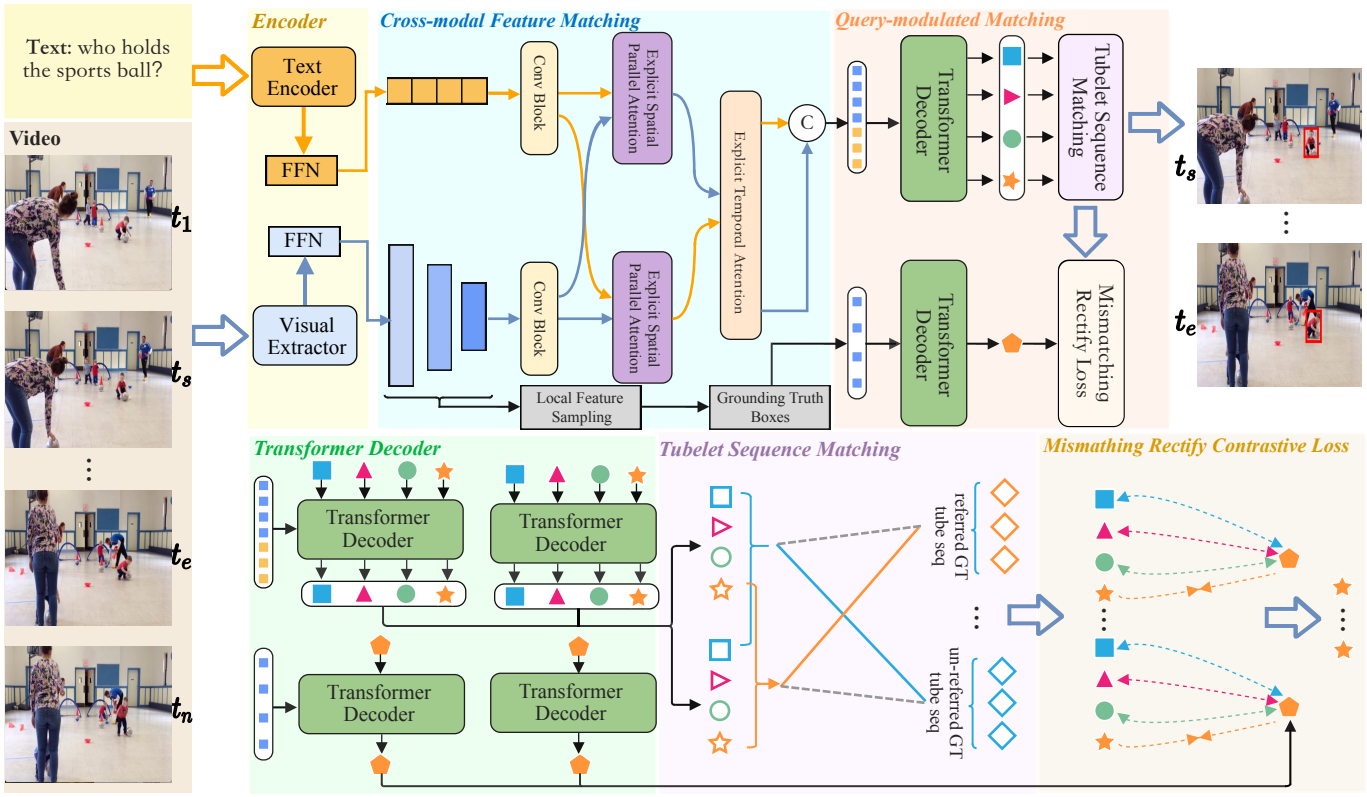


Fig. 2. An overview of the proposed Feature Semantic Matching (FSM). Given a video-query pair, FSM first encodes them and uses the cross-modal feature matching fusion module to align different modal features. The fused cross-modal features are input into the query-modulated matching module that includes: transformer decoder, tubelet sequence matching and, mismatching rectify contrastive loss module. With the transformer-decoder, N object queries and 1 object query respectively are transformed into the output embedding. Then, the output embedding is fed into the tube sequence matching module to select the concrete prediction. A new mismatching rectify contrastive loss is used to make the object query correspond to the text description meanwhile eliminating the effect of temporally disordered bounding boxes. Finally, the tube is generated based on the correct output embedding via the prediction head.

task. Inspired by the MDETR [20], TubeDETR [21] performs spatio-temporal grounding follows the one-stage framework of MDETR and reasons the temporal and spatial results. Besides, STCAT [22] proposed the spatial and temporal interaction layer to obtain better spatial-temporal features representation. However, these solutions ignored an important issue that only part of the video content is related to the text features. Previous methods lack an effective design for multi-modal feature inconsistency in STVG.

B. Video-text Interaction

Video-text tasks such as Video question answering, Video-language understanding, and Referring video object segmentation attract lots of attention for multiple-modal information processing [34], [35]. As a necessary step, visual-text interaction has an important effect on the final performance [36]–[38]. Due to the simplicity and success of Transformers in the natural language processing field, many works [33], [39]–[41] take advantage of this architecture to align the semantic information between images and text. Other works [42]–[45] further extend transformers into video-text tasks. But most of them adopt the cross-modal fusion method of static image and text, that ignored the video is continuous in time and only part of the video content is related to the text features. Considering the STVG task need to predict detailed spatial location per frame and the temporal boundary, thus these methods are not capable to handle the challenging STVG task directly.

C. Transformer based Detection

Object detection is one of the most important tasks in computer vision, early methods first use a CNN encoder or region proposal to obtain features, and then use a regression module to predict the location corresponding to the object. As *Carion et al.* [46] introduce Transformer into object detection, subsequent studies [47]–[49] have developed innovative methods leveraging Transformer architectures, applying them across various tasks, including scene recognition. However, these transformer-based detection methods focus on predicting a series of bounding boxes in the static image. Recently, the encoder-decoder paradigm [44], [45] has been used for video detection. In the encoder-decoder paradigm, each query can be regarded as a positional prior to letting decoders focus on a region of interest. TubeDETR also introduced the encoder-decoder paradigm into the spatial-temporal video grounding tasks. TubeDETR only uses a query to represent all objects which leads to inconsistencies between query and target. STCAT [22] converts multi-modal information into an object query and uses a stronger detection framework DAB-DETR to predict bounding boxes for each frame.

III. METHOD

To mitigate the aforementioned cross-modal reasoning misalignment issue and mislocalization in spatial grounding, we present the Feature Semantic Matching (FSM). The overview of FSM is illustrated in Fig. 2. FSM first utilizes two feature extractors to obtain the visual and textual features from the video and text, respectively. Then the cross-modal feature matching module (§ III-A) is designed to conduct multi-granularity alignment, which aims to align the video and text feature while enhance the spatial features and temporal features matching to the text. To further produce text-relevant object queries, the query-modulated matching module(§ III-B) is proposed to facilitate text-relevant tube construction by multiple query generation and tubulet sequence matching. Since the results produced by DETR are temporally disordered [45], the mismatching rectify contrastive loss is proposed to eliminate this effect.

Specifically, given a video-text pair input, we denote the video frame sequence as $\mathbf{V} = \{v_t\}_{t=1}^{T}$ and M words in the text as $\mathbf{Q} = \{q_m\}_{m=1}^{M}$ depicting a target object existing in $\mathbf{V}$. The goal of STVG is to localize a spatio-temporal tube $\mathbf{B} = \{b_t\}_{t=t_s}^{t_e}$ that semantically corresponds to a given text. Here b_t represents a bounding box in the t-th frame, t_s and t_e specify the starting and ending boundaries of the true object tube, respectively. For the video encoder, we use a pre-trained backbone (*e.g.*, ResNet [50]) to extract the multi-scale visual feature for each frame and flattened, bringing the sequence of the multi-scale feature $\{f^l\}_{l=1}^{L}$, where $f^l \in \mathbb{R}^{T\times H^l W^l \times d^l}$, l indexes the input feature level and d^l denotes the dimension of the l-th layer feature. For the text encoder, we leverage a pre-trained text encoder (*e.g.*, BERT [51]) to encode the text $\mathbf{Q} = \{q_m\}_{m=1}^{M}$ into its corresponding features $\{t_m\}_{m=1}^{M}$, where $t_m \in \mathbb{R}^{M\times d_t}$, and d_t is the dimension of text feature.

A. Cross-modal Feature Matching

Given the different modality features, we present a tailored transformer for effective spatio-temporal fusion. The core of the fusion lies in preserving the important spatial representation and its alignment with the text features (*i.e.*, consistency), so as to learn the better temporal expressive for temporal grounding. To this aim, we design bi-level consistency-aware feature fusion: *Frame-level features fusion* and *Video-level features fusion* as shown in Fig. 3.

Given the multi-scale feature sequence $\{f^l\}_{l=1}^{L}$ and text feature $\{t_m\}_{m=1}^{M}$, a projection layer is first applied to embed them into the same channel dimension d. We denote the projected visual embedding as $x = \{x^l\}_{l=1}^{L}$, where $x^l \in \mathbb{R}^{T\times H^l W^l \times d}$ and $y = \{y_m\}_{m=1}^{M}$, where $y_m \in \mathbb{R}^d$. Our proposed cross-modal feature fusion module takes all the above-described features as input and conducts the cross-modal features fusion layer at both *frame-level* and *video-level*.

1) Frame-level cross-modal features fusion: The STVG task requires predicting a bounding box corresponding to the text for each frame. Therefore, aligning the video and text features at the frame level and emphasizing the spatial features related to the text are beneficial for spatial grounding. Hence, the primary goal of the initial phase is to align video and textual features, while simultaneously preserving essential spatial features at the frame level through cross-modal feature fusion.

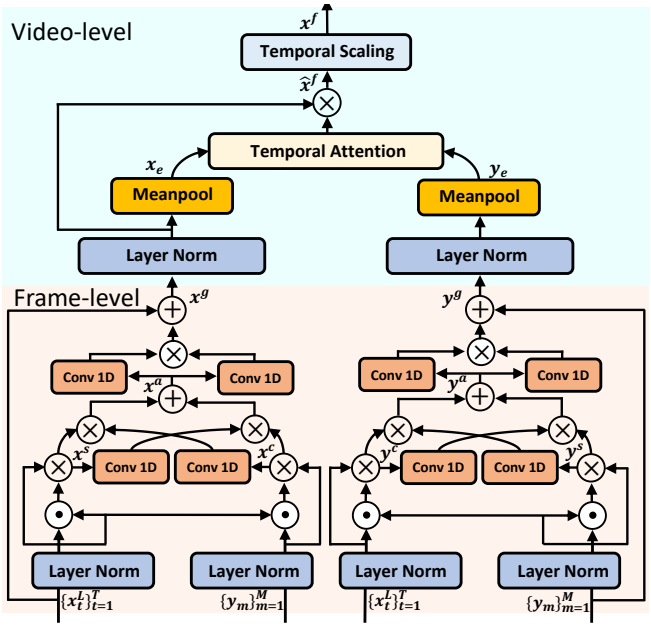


Fig. 3. The proposed cross-modal feature matching module, which performs frame and video-level cross-modal feature fusion.

The straightforward idea is to employ the transformer-style structure to model the relative transitions between x and y. However, previous transformer architecture is not suitable for STVG task. To reduce computation, we only integrate features between last level visual feature x^L and the textual features y.

In concrete, we design a *Gated Correlation* operation as shown in the lower part of Fig. 3. Specifically, the gated correlation operation is implemented to bridge the self- and cross-modal features, which can preserve important features from being suppressed.

$$\begin{aligned} x^s &= Softmax\big(x^L \odot x^L\big) \otimes x^L, \\ x^c &= Softmax\big(x^L \odot y\big) \otimes y, \end{aligned} \tag{1}$$

where $x^s = \{x_i^s\}_{i=1}^{T}$, $x^c = \{x_i^c\}_{i=1}^{T}$, $\odot$ represents the dot product, and $\otimes$ represents element-wise multiplication. Then, the self- and cross-modal representations are merged by a cross-gating:

$$x_i^a = \sigma\big(\texttt{FFN}(x_i^c)\big) \odot x_i^s + \sigma\big(\texttt{FFN}(x_i^s)\big) \odot x_i^c, \tag{2}$$

where $x^a = \{x_i^a\}_{i=1}^{T}$, σ denotes Sigmoid function, $\odot$ represents the Hadamard product, and FFN denotes the Feed-forward Network. Next, we employ a self-guided head to implicitly emphasize the informative representations by measuring the confidence of each element in x^a as:

$$x_i^g = \sigma\big(\texttt{FFN}(x_i^a)\big) \otimes \texttt{FFN}(x_i^a) + x_i^L, \quad x_i^g \in \mathbb{R}^{H^L W^L \times d}. \tag{3}$$

Finally, the frame-level cross-modal features $x^g = \{x_i^g\}_{i=1}^{T}$, where $x^g \in \mathbb{R}^{T\times H^L W^L \times d}$ is obtained. The enhanced text features $y^g = \{y_i^g\}_{i=1}^{M}$, where $y^g \in \mathbb{R}^{M\times d}$ are also obtained in a similar manner (*e.g.*, swapping video and text features).

2) Video-level cross-modal features fusion.: Considering the STVG need to localize the video clip based on the whole sentences rather than just a few words, it is meaningful to build the semantic alignment between the video and the whole sentence. Specifically, given the frame-level cross-modal features x^g and the enhanced text features y^g, we use the meanpool operation to obtain video-level spatial features $x_e \in \mathbb{R}^{T\times 1\times d}$ as well as the semantic features $y_e \in \mathbb{R}^{1\times d}$ of the entire text. Next, the visual feature x_e at video-level and semantic features y_e of the entire text are further calculated:

$$x^r = \text{Sigmoid}(\frac{x_e y_e^\top}{\sqrt{d}})x^g, \quad x^r \in \mathbb{R}^{T\times H^L W^L \times d}. \tag{4}$$

Considering sufficient temporal expressiveness is beneficial for temporal grounding [52], so we learn a set of scalars that dynamically adjust the multi-modal features x^r in Eq. 4. As shown in the upper part of Fig 3, we employ a temporal attention mechanism to implement this function. For each frame $x_t^r \in \mathbb{R}^{H^L W^L \times D}$, we impose a learnable scaling factor $\alpha_t \in \mathbb{R}^1$ with tanh-gating mechanism:

$$\begin{aligned} \alpha_t &= \tanh(\gamma_t) + 1, \quad \alpha_t \in \mathbb{R}^{T\times 1} \\ x_t^{ta} &= \alpha_t \cdot x_t^{ta}, \quad x_t^{ta} \in \mathbb{R}^{H^L W^L \times d} \end{aligned} \tag{5}$$

where $\alpha_t \in \mathbb{R}^1$ is the learnable scalar initialized at 0. The value of α_t ranges from 0 to 2. When α_t=1, our model treats each frame equally. When α_t reduces to 0, the video encoder only considers the semantics between different modalities, which does not consider any temporal dependency during the video features extraction.

Finally, we contact the fused features $x^{ta} = \{x_t^{ta}\}_{t=1}^T$ and enhanced text features y^g to obtain the enhanced multimodal features $\mathbf{F}$:

$$\mathbf{F} = concat[x^{ta}, y^g], \quad \mathbf{F} \in \mathbb{R}^{T\times(H^L W^L + M)\times d}. \tag{6}$$

B. Query-Modulated Matching

After obtaining the fused cross-modal features, the STVG task needs to predict the bounding box in each frame and associate the box to build the tube. Prior works [21] tend to utilize one query to represent all objects and transform it into output embedding via transformer decoder, leading to the misalignment between text and the detected objects easily (See Fig. 1). In order to keep the object query consistent with the target, we design a Query-Modulated Matching (QMM) module. It consists of a two-stream transformer decoder (§ III-B1), and mismatching rectify contrastive loss (§ III-B2).

1) Two-stream Transformer Decoder: Similar to DETR [20], our decoder architecture consists of the Transformer module. As shown in Fig. 2, we first feed multiply learnable queries $\{q^n\}_{n=1}^N$ and the enhanced multi-modal $\mathbf{F}$ into the upper transformer decoder $\mathcal{F}_{Tran\_dec1}$. Similar to DETR, N learnable queries are transformed into output embeddings D.

$$\begin{aligned} D &= \mathcal{F}_{Tran\_dec1}(\mathbf{F}, q^n), \quad D \in \mathbb{R}^{T\times N\times d} \\ \hat{z} &= \mathcal{F}_{seq}(D) \end{aligned} \tag{7}$$

where $\mathcal{F}_{seq}$ denotes a multiple layer prediction head. By this manner, the output embedding D from N queries are used to generate N tube $\hat{z} = \{\hat{z}^n\}_{n=1}^N$. Each $\hat{z}^n$ contains three elements: $\hat{b}_t$ means the predicted bounding boxes, $\hat{p}_t$ denotes the semantic-align probabilities, $[\hat{\tau}_t^s, \hat{\tau}_t^e]$ denotes the prediction probabilities for every frame.

As the predictions $\hat{z}$ in Eq. 7 of cross frames are disordered [45], therefore, we adopt the tubelet sequence matching to adaptively associate multiple trajectories and select the highest confident prediction to form the final tube (see Fig. 2).

In concrete, given the N queries that are used to produce N spatio-temporal trajectories $\{\hat{z}^n\}_{n=1}^N$ across multiple frames. We adopt the tubelet sequence matching to adaptively associate multiple trajectories by Hungarian algorithm [53]. The matching cost C_{cost} is computed as:

$$\begin{aligned} \hat{z}^* &= \arg\min C_{cost} \\ &= \arg\min_{\hat{z}^n\in\hat{z}} \{\lambda_{cls}\sum_{t=t_s}^{t_e}\mathcal{L}_{cls}(z_t, \hat{z}_t^n) + \lambda_{box}\sum_{t=t_s}^{t_e}\mathcal{L}_{box}(z_t, \hat{z}_t^n)\}, \end{aligned} \tag{8}$$

where $\hat{z}^* \in \hat{z}$, $z = \{p_t, b_t\}_{t=t_s}^{t_e}$ is the ground-truth tube, p_t is a one-hot label and p_t equals to 1 when z^t corresponds to the text-referred object and the ground-truth object is visible in the t-th frame, otherwise 0. b_t denotes the ground-truth bounding box. The $\mathcal{L}_{cls}(z, \hat{z}^i)$ represents the focal loss [54] while the box-related loss $\mathcal{L}_{box}$ sums up the L1 loss [21] and GIoU [21]. In this way, we can obtain the matched tube $\hat{z}^*$.

To build the semantics-aligned tube, the Query-Modulated Matching (QMM) module should learn to select the semantic-aligned box per-frame that corresponds to the language description from N candidates $\hat{z}$. Thus, we design another Transformer decoder $\mathcal{F}_{Tran\_dec2}$ that works for generating the ground-truth anchor during the model training phase:

$$x^{gt} = \mathcal{F}_{Tran\_dec2}(X_t, q^{gt}), x^{gt} \in \mathbb{R}^{(t_e - t_s + 1)\times 1\times d} \tag{9}$$

As shown in Fig. 2, the lower transformer decoder module $\mathcal{F}_{Tran\_dec2}$ is convert an ideal object feature as an anchor for contrastive loss, and the upper one is used to predict the object tube. Firstly, to produce an ideal object feature and facilitate the small object detection for the upper transformer decoder, we adopt RoIAlign [55] and multi-scale feature incorporation strategy to obtained X_t for improving the feature spatial resolution of the transformer decoder layer. We introduced the operation process in appendix VIII-A. q^{gt} denotes the learnable object query. x^{gt} corresponding to the ground-truth tube embedding with grounding $[t_s, t_e]$.

2) Mismatching rectify contrastive loss: The matching difficulty between object queries and corresponding targets is the major problem for DETR's architecture, which easily leads to mismatching. And the tube construction in Eq. 8, the tube sequence matching module may also introduce errors caused by temporally disorder queries or outliers. The errors may further cause mismatching and mislead the model training. Thus, to produce more accurate predictions and eliminate the effects of the underlying noise, we introduce a novel contrastive learning scheme to rectify the mismatching caused by the tube matching module.

In concrete, we propose a new loss function, termed mismatching rectify contrastive loss (MRCL) on the tube level. During the network training, we regard the most matched tube

$\hat{z}^*$ through the tube sequence matching as the positive sample. Thus, we use the corresponds output embedding $K^* \in \mathbb{R}^{T\times d}$ from D^n as positive embeddings while the other $N-1$ tubes output embedding in D^n are treated as negative samples $\{K^n\}_{n=1}^{N-1}$, where $K^n \in \mathbb{R}^{T\times d}$.

For anchor samples in contrastive learning, we define the output embedding x^{gt} that corresponding to the ground-truth tube are treated as the anchor sample. Imposing contrastive learning on the features of output embedding after the Transformer decoder layer can restrict the object query consistent with the target.

Thus, given the positive embedding K^*, anchor embedding x^{gt} and the negative embedding K^n, the mismatching-rectify contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{MRCL} &= -\frac{1}{\Delta \mathrm{t}} \sum\nolimits_{t=t_s}^{t_e} \log \frac{\exp(x_t^{gt}\cdot K_t^*/\tau)}{L_{pos}+L_{neg}} \quad \text{with} \\ L_{pos} &= \sum\nolimits_{t=t_s}^{t_e} \exp(x_t^{gt}\cdot K_t^*/\tau) \quad \text{and} \\ L_{neg} &= \sum\nolimits_{n=1}^{N-1}\sum\nolimits_{t=t_s}^{t_e} \exp(x_t^{gt}\cdot K_t^n/\tau), \end{aligned} \tag{10}$$

where $\Delta \mathrm{t} = t_e - t_s + 1$. L_{pos} denotes the distance between the positive sample K^* and anchor x^{gt}. L_{neg} denotes the distance between the negative sample $\{K^n\}_{n=1}^{N}$ and anchor x^{gt}. With contrastive learning loss, the distance between the positive sample K^* and the output embedding x^{gt} can be minimized, while the distance between the output embedding $\{K^n\}_{n=1}^{N-1}$ and x^{gt} is maximized. In this way, the predicted results are rectified and the impact of mismatches could be reduced.

IV. Training and Inference

1) Training: At the training stage, our model accepts video-text pairs as inputs, while each pair has a ground-truth sequence $\mathbf{B}=\{b_t\}_{t=t_s}^{t_e}$ and the corresponding start and end timestamps.

Considering STVG contains two sub-tasks: spatial localization in each frame and temporal grounding in the video. Therefore, we design two loss functions. For the spatial localization, we involve the box prediction loss and probabilities loss that all come from the cost function C_{cost} in the Hungarian algorithm Eq. 8. The only difference lies in that we only use the positive samples z^* to compute the loss. In this way, the gradient from the positive samples can facilitate the spatial localization learning.[1].

Then, we employ the $\mathcal{L}_{MRCL}$ (Eq. 10) to rectify the predicted results during tube construction. Finally, we use the tube sequence matching module to pick out positive sample z^* and follow [21] to generate two probability distribution $(\hat{\tau}^s, \hat{\tau}^e) \in [0,1]^{T\times T}$ for starting and ending positions. To encourage a more accurate temporal prediction, we also follow [21] to use a guided attention loss $\mathcal{L}_{att}$ and $\mathcal{L}_t$. L_t is the Kullback-Leibler divergence loss measuring the distance between the predicted $\hat{\tau}^e$ and the target start distribution τ^e as well as the distance between the predicted $\hat{\tau}^e$ and the target end distribution τ^e. $L_t = L_s(\hat{t}^s, t^s) + L_e(\hat{t}^e, t^e)$, where L_s and L_e are the KL divergence between target and predicted distributions. $\mathcal{L}_{att}$ is a guided attention loss that encourages weights corresponding to time queries outside of the temporal boundaries to be lower than the weights inside these boundaries. $L_{att} = -\sum_{t=t_s}^{t_e} log(1-a_i)$, a_i is the i-th column in the attention matrix A, the attention matrix A is obtained at cross-attention layer of the decoder. The total training loss is calculated as:

$$\mathcal{L} = C_{cost} + \lambda_{MRCL}\mathcal{L}_{MRCL} + \lambda_t\mathcal{L}_t + \lambda_{att}\mathcal{L}_{att}. \tag{11}$$

2) Inference: During inference, our FSM will predict N tube sequence by giving the video-text pairs. For each tube query sequence, we average the predicted probabilities over all the frames and obtain the score set:

$$\hat{P}^n = mean(\hat{p}_t), t \in [1, ...T]. \tag{12}$$

Then, we select the optimal tube sequence with the highest average score and obtain its index σ as:

$$\sigma = \underset{n\in\{1,2,...,N\}}{\arg\max}\ \hat{P}^n. \tag{13}$$

After obtaining index σ, we select the corresponding feature embedding from D^n. Next, we use the D^n to obtain the spatial-temporal tube $\hat{z}^\sigma$. The final predicted object bounding box $\hat{b}_t$, semantic-align probabilities $\hat{p}_t$, and prediction probabilities $[\hat{\tau}_t^s, \hat{\tau}_t^e]$ for every frame are obtained by MLP prediction head. The start and end times of the output tube, denoted as t_s and t_e, are computed by selecting the maximum of the joint start and end probability distribution $(\hat{\tau}^s, \hat{\tau}^e) \in [0,1]^{T\times T}$, while invalid combinations where $t_s \le t_e$ are masked out.

V. Experiments

A. Datasets and Metrics

1) Datasets: To evaluate the proposed method, we follow previous work [21] and adopt two large video grounding datasets: VidSTG [25] and HC-STVG [26].

- **VidSTG** consists of 99,943 sentences with 44,808 declarative sentences and 55,135 interrogative sentences describing 80 types of objects appearing in 10,303 videos. The dataset is divided into training, validation and, test with 80,684/ 8,956/ 10,303 distinct sentences respectively, and 5,436/ 602/ 732 distinct videos respectively.
- **HC-STVG** consists of videos in multi-person scenes, each annotated with one sentence referring to a person. This dataset is divided into training and validation subsets with 10,131 and 2,000 video-sentence paris.

2) Evaluation Metrics: Following the previous works [21], we utilize the **m_vIoU**, **m_tIoU** and **vIoU@R** as the evaluation metric. The $vIoU = \frac{1}{|S_u|}\sum_{t\in S_i} \mathrm{IoU}(\hat{b}_t, b_t)$, where S_i and S_u are the intersection and union (IoU) between the predicted tubes and ground-truth tubes, respectively. Concretely, we first compute the IoU score between the predicted bounding box $\hat{b}_t$ and ground truth b_t at frame t. The m_vIoU score is defined as the average $vIoU$ score over all testing videos. The m_vIoU is the ratio of samples $vIoU > R$. To evaluate spatial grounding only, we use the m_tIoU (tIoU = $\frac{|S_i|}{|S_u|}$), which is computed by using GT start and end times.

[1]Note that the positive samples are used to predict the ground-truth labels and bounding boxes, while the negative samples are used to predict the ∅ label instead.

TABLE I
PERFORMANCE COMPARISONS OF THE STATE-OF-THE-ARTS ON THE VIDSTG TEST SET(%).

Methods	Resolution	Parameters	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 [56]+TALL [57]	-	-	-	9.78	11.04	4.09	-	9.32	11.39	3.24
STPR [15]+TALL [57]	-	-	34.63	10.40	12.38	4.27	33.73	9.98	11.74	4.36
WSSTG [11]+TALL [57]	-	-	-	11.36	14.63	5.91	-	10.65	13.90	5.32
GroundeR [56]+L-Net [58]	-	-	-	11.89	15.32	5.45	-	11.05	14.28	5.11
STPR [15]+L-Net [58]	-	-	40.86	12.93	16.27	5.68	39.79	11.94	14.73	5.27
WSSTG [11]+L-Net [58]	-	-	-	14.45	18.00	7.89	-	13.36	17.39	7.06
Two-Stage:										
STGRN [25]	-	-	48.47	19.75	25.77	14.60	46.98	18.32	21.10	12.83
STGVT [26]	-	-	-	21.62	29.80	18.94	-	-	-	-
OMRN [59]	-	-	50.73	23.11	32.61	16.42	49.19	20.63	28.35	14.11
One-Stage:										
STVGBert [23]	-	-	-	23.97	30.91	18.39	-	22.51	25.97	15.95
TubeDETR [21] [one query]	224	65.63M	46.90	27.60	37.70	25.70	46.10	23.30	31.30	20.80
STCAT [22] [one query]	224	87.14M	<u>48.76</u>	<u>28.04</u>	<u>39.44</u>	<u>26.00</u>	<u>47.50</u>	<u>23.32</u>	<u>32.26</u>	<u>21.27</u>
FSM (Ours) [multi query]	224	67.61M	**49.09**	**28.23**	**39.52**	**26.95**	**47.56**	**23.43**	**32.54**	**21.53**

The number of parameters does not include the number of parameters for the text encoder.

TABLE II
PERFORMANCE COMPARISONS OF THE STATE-OF-THE-ARTS ON THE HC-STVG TEST SET(%).

Methods	Resolution	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Two-Stage:					
STGVT [26]	-	-	18.15	26.81	9.48
One-Stage:					
STVGBert [23]	-	-	20.42	29.37	11.31
TubeDETR [21]	352	43.70	32.40	49.80	23.50
STCAT [22]	224	**47.18**	<u>31.79</u>	<u>51.35</u>	<u>26.97</u>
FSM (Ours)	224	<u>47.15</u>	**34.00**	**54.14**	**29.74**

3) Implementation Details: The output embedding of the upper transformer decoder is used for visual grounding and temporal localization to simultaneously obtain predictions for *all frames of the video*. The refined output embedding D^m from N queries are used to generate N tube $\hat{z} = \{\hat{z}^i\}_{i=1}^N$. Every tube contains three components: the predicted bounding boxes, the visual grounding (start and end times), and the probabilities of text-referred object. In detail, normalized coordinates of all bounding boxes (2D center and size) $\hat{b} \in [0,1]^{T\times 4}$ are predicted by a 3-layer MLP. Probabilities of the video tube, $\hat{\tau}_t^s \in [0,1]^T$ and $\hat{\tau}_t^e \in [0,1]^T$ are predicted with 2-layer MLPs. The probabilities of the text-referred object $\hat{p} \in [0,1]^T$ are the probability scalar indicating whether the bounding box corresponds to the text-referred object and the object is visible in the current frame. The probabilities of text-referred object are predicted with a linear layer. Here, for the prediction results $\hat{z}^t = \{\hat{p}_t^i, \hat{b}_t^i, \hat{\tau}_t^s, \hat{\tau}_t^e\}_{t=1}^T$ are obtained for t-th frmae.

In line with the previous methods, we adopt the ResNet-101 [50] as the visual encoder and RoBERTa [60] as the linguistic encoder. For the encoder and decoder, the number of attention heads is set to 8 and the hidden dimension of feed-forward networks in the attention layer is 2048.

Following previous methods [21], [22], the model parameters are firstly initialized with the pre-trained weights provided in [20] and then the whole framework is end-to-end optimized during model training. We use the TubeDETR [21] which needs 16 V100 as the baseline. Due to the appearance similarity between adjacent frames, we tune the uniform down-sample from 200 to 100 on TubeDETR [21] to reduce the computation cost. We also use same data augmentations with TubeDETR including random resizing and random cropping to all training videos. The final object tube is obtained by linearly interpolating the predicted bounding box in sampled frames.

During the training process, $\mathcal{F}_{Tran\_dec2}$ and $\mathcal{F}_{Tran\_dec2}$ share parameters. We empirically set the other hyper-parameters $\lambda_{cls} = 0.2$, $\lambda_{MRCL} = 2$ in VidSTG and $\lambda_{cls} = 0.2$, $\lambda_{MRCL} = 1$ in HC-STVG. Follow previous method [21], we set $\lambda_{box} = 1$, $\lambda_t = 10$, and $\lambda_{att} = 1$. We use AdamW optimizer with weight-decay $1e^{-4}$ and initial learning rates $1e^{-5}$ for the visual backbone, and $5e^{-5}$ for the rest of the network.

B. Performance Comparison

To verify the effectiveness of FSM, we perform the comparisons with representative models on VidSTG and HC-STVG in Table I and II. Note that the best results are highlighted in **bold** while the second-best is <u>underlined</u>. Methods for the STVG task are divided into two types, namely **Two-Stage** and **One-Stage**. The two-stage approaches consists of the STGRN [25], STGVT [26], and OMRN [59]. These methods first generate box proposals in each frame by a pre-trained object detector and then select the best matches from these candidates to accomplish spatio-temporal grounding. The one-stage methods like STVGBert [23] and recent concurrent works TubeDETR [21], STCAT [22] by a unified architecture to predict the spatio-temporal tube from the given video-text pair.

The STVG task needs a high amount of computing resources (*e.g.*, TubeDETR uses 16 V100 and STCAT uses 32 A100). Due to the limited computing resources, we have adjusted the hyper-parameters such as resolution and the number of frames. Our

method utilizes TubeDETR [21] as the backbone architecture. From the source code of STCAT [22] and TubeDETR [21], we can find that STCAT employs more robust data augmentation techniques by using the greater resolution in the training and inference stage. In order to achieve a fair comparison, we adjust the resolution and data augmentation to keep the same settings as TubeDETR.

Table II demonstrates the detailed results comparison of the proposed model. Through the analysis of the experimental results, we further obtained the following observations. 1) The proposed cross-modal features fusion considers self- and cross-modal features between words, text, and video frames, which better aligns the feature between video and text. Meanwhile, this fusion strategy preserves the more useful features in terms of space and time. This is more suitable for this task than other methods that simply concatenate video text features or use the attention mechanism. 2) The two-stage approach leads to the domain gap imposed by the pre-trained detectors. Our one-stage methods can break the restriction of the domain gap through end-to-end training. 3) Our proposed method uses multiple queries and selects the relatively concrete result by Hungarian matching algorithm. Thus, our method yields better performance than TubeDETR (49.09 versus 46.90, 47.56 versus 46.10). Next, compared to the one-stage methods, the query-modulated matching module constrains the query to the spatial area corresponding to the text information. In this way, the query in our method is sensitive to text information and makes the predicted bounding boxes more consistent with the ground-truth bounding boxes.

As the main counterpart, STCAT [22] solves the problem of text-agnostic object query by converting multi-modal information into an object query. To obtain more accurate prediction results, STCAT uses a stronger detection framework DAB-DETR and uses different transformer decoder modules to predict spatial and temporal results separately. Therefore, STCAT model has a huge amount of parameters for the visual module. Our FSM with multiple learnable queries and query-modulated matching module achieves comparable results with STCAT but with fewer parameters than STCAT, especially for the HC-STVG dataset (34.00 versus 31.79, 54.14 versus 51.35, 29.74 versus 26.97).

C. Ablation Study

In this section, we conduct some ablations on the HC-STVG benchmark to further investigate the contributions of different components in the proposed framework. To reduce computational source, we only conduct ablation study at a resolution of 224 on the HC-STVG dataset, which only need 8 RTX3090 GPUs.

TABLE III
EFFECT OF THE CROSS-MODAL FEATURE MATCHING MODULE.

frame	video	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
✘	✘	44.52	31.43	50.60	27.07
✓	✘	45.84	32.59	53.62	27.93
✘	✓	46.10	32.72	51.64	27.67
✓	✓	**47.15**	**34.00**	**54.14**	**29.74**

TABLE IV
EFFECT OF THE GATE OPERATION.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
with gate	**47.15**	**34.00**	**54.14**	**29.74**
without gate	44.59	31.14	48.02	27.50

1) Effect of the cross-modal feature matching module: To demonstrate the validation of proposed frame- and video-level, we compare our full model with variants removing cross-modal feature matching module. The quantitative ablation results are shown in Table III. We find that video-level cross-modal fusion is more beneficial for the temporal grounding performance (+1.58% on m_tIoU). Furthermore, we can observe the frame-level cross-modal fusion brings a huge gain for spatial grounding (+1.16% on m_vIoU, +0.86% on vIoU@0.5). Finally, the spatio-temporal achieved significant improvement by using both frame- and video-level cross-modal features fusion.

TABLE V
EFFECT OF THE QUERY-MODULATED MATCHING (QMM).

Setting	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
with QMM	43.17	31.35	48.97	24.14
without QMM	**47.15**	**34.00**	**54.14**	**29.74**

TABLE VI
EFFECT OF THE NUMBER OF QUERIES.

Number of query	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
1	45.39	32.67	{53.17	28.36
5	**47.15**	**34.00**	**54.14**	**29.74**
10	46.54	33.41	53.36	29.14
15	46.89	33.59	53.79	28.62
25	46.18	32.91	52.41	28.45

TABLE VII
THE PARAMETER QUANTITY FOR EACH MODULE.

frame-level	video-level	contrative loss	Params
✘	✘	✘	185.63M (TubeDETR)
✓	✘	✘	187.21M
✘	✓	✘	186.03M
✘	✘	✓	185.63M

2) Effect of the gate operation: We operate the ablation study to further verify the validity of the gate operation as shown in Table IV. We can observe the gate operation bring a huge gain for spatial-temporal video grounding (+2.56% on m_tIoU, +2.86% on m_vIoU, +6.12% on vIoU@0.3, and +2.24%vIoU@0.5). Therefore, the specially designed gate operation is effective for spatial-temporal grounding.

3) Effect of the query-modulated matching (QMM): The query-modulated matching and tube sequence matching ensure the predicted bounding boxes be more consistent with the ground-truth bounding box. We compare one query (used in TubeDETR [21]) with multiple queries to explore the function of the query-modulated matching module. The detailed ablation results are shown in Table V. From the comparison, we can observe a distinct performance promotion with the proposed module (+3.98% on m_tIoU, +5.17% on vIoU@0.3, +5.60%

Fig. 4. Some visualization examples of the spatio-temporal video grounding predictions produced by the TubeDETR (blue) and our model (yellow), compared with annotated ground truth (green) on VidSTG dataset.

TABLE VIII
EFFECT OF SHARING PARAMETERS FOR TWO DECODERS.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
shared weights	**47.15**	**34.00**	**54.14**	**29.74**
unshared weights	46.46	33.37	52.67	26.64

TABLE IX
EFFECT OF THE SCALING FACTOR.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
with scaling factor	**47.15**	**34.00**	**54.14**	**29.74**
without scaling factor	46.25	33.52	53.22	27.67

TABLE X
EFFECT OF THE RESOLUTION.

Resolution	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
224	47.15	34.00	54.14	29.74
320	47.92	34.69	55.12	30.17
352	48.63	35.06	55.48	30.61

on vIoU@0.5), which further validates the effectiveness of our proposed query-modulated matching module.

4) Effect of the number of queries: From the below table VI, we can find that the performance of using multiple queries is significantly better than one query. The model performance first rises along with the increase of the object queries, then it drops when the number of object queries continuing increase as shown in Tab. VI. Thus, there is a trade-off between model performance and the number of object queries. Based on this, we have decided to set the optimal number of queries at 5.

5) Effect of sharing parameters for two decoders: We also conduct the experiment to verify the importance of sharing

parameters of two decoders. As shown in Table VIII, using different parameters would harm the performance.

6) The parameter quantity for each module: We operate the ablation study to analysis about the parameter of each component in the proposed method. The backbone contains 185.63M parameters, while our proposed frame-level, video-level, and contrastive leaning modules only contain 1.58M, 0.39M parameters, respectively. During model training, we use share parameters for $\mathcal{F}_{Tran\_dec1}$ and $\mathcal{F}_{Tran\_dec2}$. Therefore, the cotrastive module generates almost no new parameters. Compared to the backbone, our added parameters only occupy a small portion (1.1%) among the whole parameters.

7) Effect of the scaling factor: The learnable scaling factor can further enhance the output of Eq. 4. As shown in Table IX, the scaling factor can further improve the performance.

8) Effect of the resolution: From Table X, we can find that as the resolution increases, the performance of the model gradually increases. This observation aligns with findings from STCAT [22], which noted that increasing the input frame resolution leads to an enhancement in performance on the STVG dataset. This improvement can be attributed to the fact that a larger spatial scale provides more fine-grained visual clues for multi-modal reasoning. Consequently, it is evident that the resolution of the input video significantly influences grounding performance.

9) Qualitative analysis: In this section, we illustrate some examples in Fig. 4 to qualitatively compare our method with the TubeDETR [21] on the VidSTG dataset. As shown in Fig. 4, the GT (green) represents the ground-truth boxes and the corresponding temporal boundary. We compare the TubeDETR (blue) with our method (yellow). As shown in this figure, our method can generate closer boundaries as well as more accurate bounding boxes compared with TubeDETR. The last row shows failure cases. The reason is that there are two "balls" in the video and the target "ball" (green) is very small, so FSM mistakenly grounds the other "ball" as the target.

Limitations The STVG task requires huge computing resources due to the videos of the VidSTG dataset are relatively complex and the quality is relatively poor. Therefore, the larger resolution is better for VidSTG. Thus, we can extract video features and text features through pre-training models and save them to reduce the computational source in future works. In addition, the performance of the model is limited by the pre-trained model. Using a more advanced pre-trained model can yield better performance.

VI. Conclusion

In this paper, we propose a semantic-aligned matching network to address the spatio-temporal video grounding task. To address feature misalignment issues between video and text in existing cross-modal feature fusion methods, we propose a specially designed consistency-aware feature fusion module to operate frame-level and video-level features fusion to preserve the spatial representation and better alignment of video features with text features. In order to generate the text-relevant object query and to decode the desired object for the corresponding text, we design the query-modulate restriction module. We use multiple queries and convert the multi-modal feature and multiple queries to output embedding. We adopt the tubelet sequence matching to associate multiple trajectories and select the highest confidence. Finally, we restrict and rectify the query to eliminate the effects of underlying noise by mismathing rectify contrastive learning. Experiments show that our method outperforms the state-of-the-art methods by large margins on both VidSTG and HC-STVG.

VII. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 62176139, 62106128, 62176141), the Major basic research project of Shandong Natural Science Foundation (No. ZR2021ZD15), the Natural Science Foundation of Shandong Province (No. ZR2021QF001), the Young Elite Scientists Sponsorship Program by CAST (No. 2021QNRC001), the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), Shandong Province Science and Technology Small and Medium-sized Enterprise Innovation Capacity Enhancement Project (2023TSGC0115), Shandong Province Higher Education Institutions Youth Entrepreneurship and Technology Support Program (2023KJ027), the Taishan Scholar Project of Shandong Province (tsqn202103088), the Open Research Project Programme of the State Key Laboratory of Internet of Things for Smart City (University of Macau) (Ref. No.:SKL-IoTSC(UM)-2021-2023/ORP/GA05/2022).

References

[1] Cheng, Zhi-Qi and Wu, Xiao and Liu, Yang and Hua, Xian-Sheng, "Video2shop: Exact matching clothes in videos to online shopping images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4048–4056.

[2] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua, "Video ecommerce++: Toward large scale online video advertising," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1170–1183, 2017.

[3] Z.-Q. Cheng, H. Zhang, X. Wu, and C.-W. Ngo, "On the selection of anchors and targets for video hyperlinking," in *Proceedings of the 2017 acm on international conference on multimedia retrieval*, 2017, pp. 287–293.

[4] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[5] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 15–24.

[6] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, and L. S. Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257.

[7] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng, W. Wei, Z. Xu, and Y. Xie, "Context-aware biaffine localizing network for temporal sentence grounding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 235–11 244.

[8] D. Liu, X. Qu, and W. Hu, "Reducing the vision and language bias for temporal sentence grounding," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4092–4101.

[9] D. Liu, X. Fang, P. Zhou, X. Di, W. Lu, and Y. Cheng, "Hypotheses tree building for one-shot temporal sentence localization," *arXiv preprint arXiv:2301.01871*, 2023.

[10] Z. Li, Y. Guo, K. Wang, F. Liu, L. Nie, and M. Kankanhalli, "Learning to agree on vision attention for visual commonsense reasoning," *IEEE Transactions on Multimedia*, 2023.

[11] Z. Chen, L. Ma, W. Luo, and K.-Y. K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," in *ACL*, 2019.

[12] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6543–6554.

[13] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Parallel attention network with sequence matching for video grounding," *arXiv preprint arXiv:2105.08481*, 2021.

[14] X. Lu, L. Zhu, Z. Cheng, J. Li, X. Nie, and H. Zhang, "Flexible online multi-modal hashing for large-scale multimedia retrieval," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1129–1137.

[15] M. Yamaguchi, K. Saito, Y. Ushiku, and T. Harada, "Spatio-temporal person retrieval via natural language queries," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1453–1462.

[16] A. B. Vasudevan, D. Dai, and L. Van Gool, "Object referring in videos with language and human gaze," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4129–4138.

[17] X. Sun, J. Gao, Y. Zhu, X. Wang, and X. Zhou, "Video moment retrieval via comprehensive relation-aware network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[18] J. Gao and C. Xu, "Learning video moment retrieval without a single annotated video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1646–1657, 2022.

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

[20] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[21] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Tubedetr: Spatio-temporal video grounding with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 442–16 453.

[22] Y. Jin, Z. Yuan, Y. Mu *et al.*, "Embracing consistency: A one-stage approach for spatio-temporal video grounding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 192–29 204, 2022.

[23] R. Su, Q. Yu, and D. Xu, "Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1533–1542.

[24] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, and J. Wang, "Group detr: Fast detr training with group-wise one-to-many assignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6633–6642.

[25] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. 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*, 2020, pp. 10 668–10 677.

[26] Z. Tang, Y. Liao, S. Liu, G. Li, X. Jin, H. Jiang, Q. Yu, and D. Xu, "Human-centric spatio-temporal video grounding with visual transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8238–8249, 2021.

[27] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 843–851.

[28] G. Gong, L. Zhu, and Y. Mu, "Language-guided multi-granularity context aggregation for temporal sentence grounding," *IEEE Transactions on Multimedia*, 2022.

[29] Z. Xu, K. Wei, X. Yang, and C. Deng, "Point-supervised video temporal grounding," *IEEE Transactions on Multimedia*, 2022.

[30] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 870–12 877.

[31] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. v. d. 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*, 2019, pp. 1960–1968.

[32] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, "Improving referring expression grounding with cross-modal attention-guided erasing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1950–1959.

[33] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[34] J. Sun, F. Xue, J. Li, L. Zhu, H. Zhang, and J. Zhang, "Tsinit: a two-stage inpainting network for incomplete text," *IEEE Transactions on Multimedia*, 2022.

[35] X. Wang, Z. Zheng, Y. He, F. Yan, Z. Zeng, and Y. Yang, "Progressive local filter pruning for image retrieval acceleration," *IEEE Transactions on Multimedia*, 2023.

[36] X. Sun, X. Wang, J. Gao, Q. Liu, and X. Zhou, "You need to read again: Multi-granularity perception network for moment retrieval in videos," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1022–1032.

[37] J. Gao, X. Sun, M. Xu, X. Zhou, and B. Ghanem, "Relation-aware video reading comprehension for temporal language grounding," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3978–3988, 2021.

[38] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Temporal sentence grounding in videos: A survey and future directions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[40] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2020.

[41] X. Wang, L. Zhu, Z. Zheng, M. Xu, and Y. Yang, "Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision," *IEEE Transactions on Multimedia*, 2022.

[42] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," in *EMNLP*, 2020.

[43] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8746–8755.

[44] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4985–4995.

[45] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4974–4984.

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

[47] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[48] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022.

[49] Z.-Q. Cheng, Q. Dai, S. Li, T. Mitamura, and A. Hauptmann, "Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3272–3281.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[51] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[52] D. Chen, C. Tao, L. Hou, L. Shang, X. Jiang, and Q. Liu, "Litevl: Efficient video-language learning with enhanced spatial-temporal modeling," *arXiv preprint arXiv:2210.11929*, 2022.

[53] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8741–8750.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[56] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *European Conference on Computer Vision*, 2016, pp. 817–834.

[57] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.
[58] J. Chen, L. Ma, X. Chen, Z. Jie, and J. Luo, "Localizing natural language in videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8175–8182.
[59] Z. Zhang, Z. Zhao, Z. Lin, B. Huai, and N. J. Yuan, "Object-aware multi-branch relation networks for spatio-temporal video grounding," *IJCAI*, 2020.
[60] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

VIII. Appendix

In the supplement file, we present more details of RoIAlign operation [55].

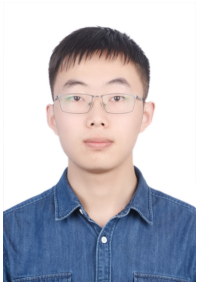
A. RoIAlign and multi-scale feature incorporation strategy

For the lower transformer decoder $\mathcal{F}_{Tran\_dec2}$, we use the enhanced multi-modal features as input. By the RoIAlign operation, our model obtains the multi-scale local features under the guidance of the ground-truth bounding boxes as :

$$\begin{aligned} X_t^l &= \text{RoIAlign}(x_t^l, b_t), \quad t_s \le t \le t_e \\ X_t &= [X_t^1, ..., X_t^l, ..., X_t^L], \quad t_s \le t \le t_e \end{aligned} \tag{14}$$

where b_t is the ground-truth bounding boxes, $X_t^l \in \mathbb{R}^{S^2 \times d}$ is the l-th layer sampled feature, S is the feature resolution in RoIAlign sampling [55]. $X_t \in \mathbb{R}^{LS^2 \times d}$ is the multi-scale sample feature by concat the sample feature from 1-th layer to L-th layer.

The multi-scale sample features X_t comes from the ROIAlign modules applied on the visual embedding x.



Tong Zhang is a master student in the School of Software, Shandong University. He received the bachelor degree from Shandong Agricultural University in 2021. His research interests including computer vision, deep learning, temporal localization with natural language.

Hao Fang Hao Fang is currently pursuing the M.S. degree with the School of software, Shandong University, Jinan, China. He received the B.S. degree in intelligence science and technology from Hangzhou Dianzi University, Hangzhou, China, in 2022. His research interests include computer vision, saliency detection, and video segmentation.

Hao Zhang is received his PhD degree in computer science from Nanyang Technological University, Singapore, in 2022. He has published multiple journal and conference papers in IEEE TPAMI, CVPR, ACL, EMNLP, SIGIR, AAAI, etc. His research areas include vision-language learning, natural language processing and interactive recommentation, conversational recommendation, large language models, etc.

Jialin Gao received the Ph.D. degree from the School of Electronic Information and Electrical Engineering of Shanghai Jiao Tong University, in 2022. Prior to that, he received the B.S. degree in electronic information engineering from the University of Electronic Science and Technology of China, in 2016. He is currently a research fellow of AI Singapore, National University of Singapore. His research interests include but are not limited to human-centeric computer vision, vision-and-language understanding, and multi-modal learning, large language models.

Xiankai Lu is a research Professor in the School of Software, Shandong University. From 2018 to 2020, he was a research associate with Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He received the Ph.D. degree from Shanghai Jiao Tong University in 2018. His research interests include computer vision, object tracking, video object segmentation and deep learning.

Xiushan Nie received the PhD degree from Shandong University, Jinan, China, in 2011. He is a professor with Shandong Jianzhu University, Jinan, China. From 2013 to 2014, he was a visiting scholar at the University of Missouri-Columbia. His research interests include data mining, multimedia retrieval and indexing, and computer vision. He is a member of the IEEE.

Yilong Yin received the Ph.D. degree from Jilin University, Changchun, China, in 2000. From 2000 to 2002, he was a Post-Doctoral Fellow with the Department of Electronics Science and Engineering, Nanjing University, Nanjing, China. He is currently the Director of the Data Mining, Machine Learning, and their Applications Group and a Professor of the School of Software Engineering, Shandong University, Jinan, China. His research interests include machine learning, data mining,and computational medicine.