Collaborative Static and Dynamic Vision-Language Streams for Spatio-Temporal Video Grounding

Zihang Lin[1], Chaolei Tan[1], Jian-Fang Hu[1,3,4]*, Zhi Jin[1], Tiancai Ye[2], Wei-Shi Zheng[1,3,4]
[1]Sun Yat-sen University, China [2]Tencent, China
[3]Guangdong Province Key Laboratory of Information Security Technology, China
[4] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{linzh59, tanchlei}@mail2.sysu.edu.cn, {hujf5, jinzh26}@mail.sysu.edu.cn,
tiancaiye@tencent.com, wszheng@ieee.org

Abstract

Spatio-Temporal Video Grounding (STVG) aims to localize the target object spatially and temporally according to the given language query. It is a challenging task in which the model should well understand dynamic visual cues (e.g., motions) and static visual cues (e.g., object appearances) in the language description, which requires effective joint modeling of spatio-temporal visual-linguistic dependencies. In this work, we propose a novel framework in which a static vision-language stream and a dynamic vision-language stream are developed to collaboratively reason the target tube. The static stream performs cross-modal understanding in a single frame and learns to attend to the target object spatially according to intra-frame visual cues like object appearances. The dynamic stream models visual-linguistic dependencies across multiple consecutive frames to capture dynamic cues like motions. We further design a novel cross-stream collaborative block between the two streams, which enables the static and dynamic streams to transfer useful and complementary information from each other to achieve collaborative reasoning. Experimental results show the effectiveness of the collaboration of the two streams and our overall framework achieves new state-of-the-art performance on both HCSTVG and VidSTG datasets.

1. Introduction

Vision-language cross-modal understanding is a challenging yet important research problem that bridges the communication between humans and artificial intelligence systems. It has attracted increasing attention and many vision-language tasks were studied in recent years, like visual grounding [13, 32], VQA [7, 31], image/video captioning [15, 40], etc. In this work, we focus on a challenging vision-language task named Spatio-Temporal Video Grounding (STVG) which was recently proposed in [42]. Given a language query indicating an object (as shown in Figure 1), STVG aims to localize the target object spatially and temporally in the video. In this task, the input language query may express different kinds of visual concepts, thus the model requires to well capture and understand these concepts in both vision and language modalities.

In STVG task, dynamic visual concepts like human motion and static visual concepts like object appearance are both important for distinguishing the target object from other objects that occurred in the same video. For example, in the first sample in Figure 1, the two men are dressed alike (i.e., their static appearance cues are similar), and we can only distinguish them by motion. In the second sample, the two women perform similar actions, they both stand up (i.e., they have similar dynamic motion cues), here, we can only distinguish them by their clothes. The above examples show that static or dynamic visual cues alone cannot solve the STVG task well. And it implies that modeling static and dynamic visual-linguistic dependencies and collaboratively utilizing them are important for addressing STVG task.

Humans treat static and dynamic cues differently [4, 21], but this was overlooked in previous STVG works [22, 24, 30]. Taking the first query in Figure 1 as an example, a human would randomly or evenly click on some locations on the video's progress bar to find candidate frames containing a man in blue coat. And he will play the video around that frame and **attend** on the candidate man to **check** whether he performs the action described in the text (i.e., "turns around and stops by the stone"). In the above process, the human understands static and dynamic cues in different ways (i.e., view a single frame and watch the video clip, respectively) and determines the target object by jointly considering the

*Corresponding author.

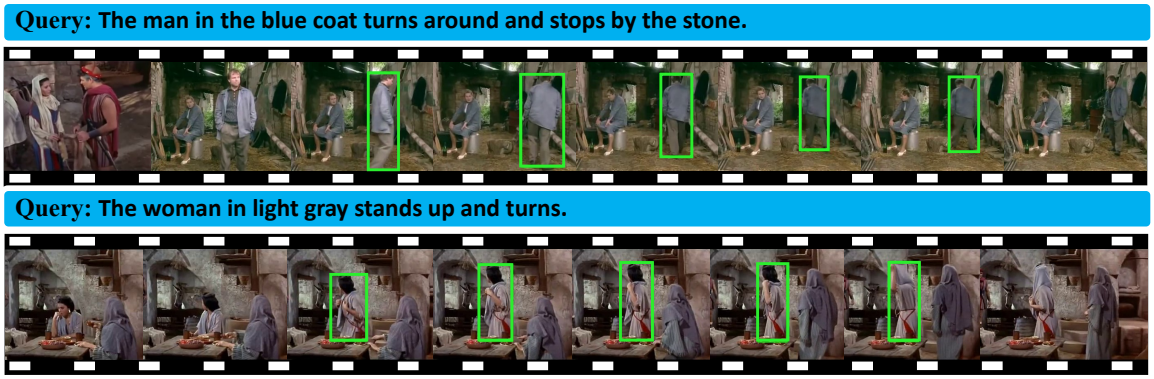

Figure 1. Two examples for Spatio-Temporal Video Grounding task. In the first case, the two men are both dressed alike in blue, thus understanding the action described in the query sentence is essential to recognize the person of interest. In the second case, both of the two women stand up, we can only distinguish them by their clothes. (Best viewed zoomed in on screen.)

static and dynamic cues in an "attend-and-check" process. Inspired by the above observations, we propose a framework that consists of a static and a dynamic vision-language (VL) streams to model static and dynamic cues, respectively. And we design a novel cross-stream collaboration block between the two streams to simulate the "attend-and-check" process, which exchanges useful and complementary information learned in each stream and enables the two streams to collaboratively reason the target object.

Specifically, in this work, our static vision-language (VL) stream learns to attend to some candidate regions according to static visual cues like appearance, while the dynamic VL stream learns to understand the dynamic visual cues like action described in the text query. Then, in the collaboration block, we guide the dynamic stream to only focus and *attend* on the motion of the candidate objects by using the learned attended region in the static stream. And we transfer the text-motion matching information learned in the dynamic stream to the static stream, to help it further *check* and determine the target object and predict a more consistent tube. With the above cross-stream collaboration blocks, both the static and dynamic vision-language streams can learn reciprocal information from the other stream, which is effective for achieving more accurate spatio-temporal grounding predictions. We conduct experiments on HCSTVG [24] and VidSTG [42] datasets and our approach outperforms previous approaches by a considerable margin. Ablation studies demonstrate the effectiveness of each component in our proposed cross-stream collaboration block and show its superiority over commonly used counterparts in video understanding works [4, 21].

In summary, our contributions are: 1), we develop an effective framework that contains two parallel streams to model static-dynamic visual-linguistic dependencies for complete cross-modal understanding; 2), we propose a novel cross-stream collaboration block between the two streams to exchange reciprocal information for each other and enable collaborative reasoning of the target object; 3), Our overall framework achieves state-of-the-art performances on HCSTVG [24] and VidSTG datasets [42].

2. Related Work

Spatio-Temporal Video Grounding. Spatio-temporal video grounding aims to localize the target object spatially and temporally according to a language query. Early works [24, 41, 42] employ pre-trained detectors like Faster R-CNN [18] to detect objects in each frame, and build their model upon the detection features of each object. STGRN [42] and OMRN [41] learn object relations with spatio-temporal graphs and multi-branch relation networks, respectively. STGVT [24] learns cross-modal representations with visual transformers for video-sentence matching and temporal localization. However, these methods are limited to the pre-detected bounding boxes and cannot localize object categories that are not defined in the pre-training dataset. Recent works [22, 30] design their framework based on strong pre-trained vision-language models and train their models end-to-end. These approaches do not require pre-detected bounding boxes and they perform much better. However, they directly collapse some dimensions (i.e., spatial or temporal) in cross-modal attention calculation to reach an acceptable computation cost, which results in large information loss, and thus they can not well capture the important static-dynamic cross-modal contextual relationship. To address this weakness, we propose to model static and dynamic visual-linguistic correspondence with parallel streams and learn to reason the target object in a collaborative framework.

Temporal Grounding. Given a language query, temporal grounding aims to predict the time span in a video specified by the language query [1, 6]. Existing works [11, 16, 17, 29, 35, 36, 38, 39] can be divided into proposal-based and proposal-free approaches. Proposal-based approaches generate some proposals (manually designed [29, 39] or

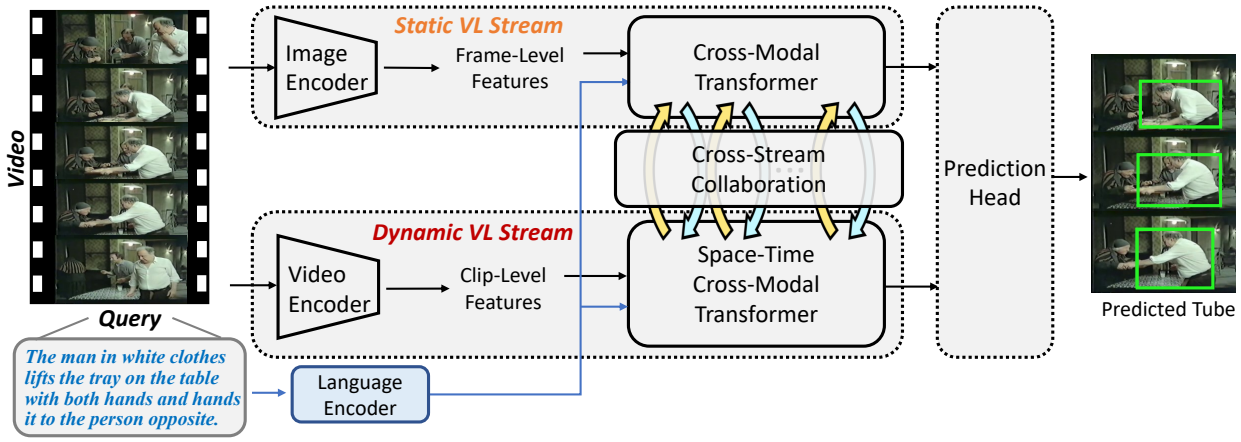

Figure 2. An overview of the proposed framework. Our framework mainly consists of a static VL(vision-language) stream and a dynamic VL stream. The static stream learns to attend to the spatial locations of the target object according to static cues like object appearance. The dynamic stream learns motion-text correspondence according to dynamic cues like human action. We further devise a cross-stream collaboration block that enables the two streams to query useful and complementary information from the other stream.

predicted by the model [35]) and predict the IoU scores between each proposal and the ground truth. Proposal-free approaches address moment localization by directly regressing the starting/ending points of the target time spans [11, 16] or predicting the probability of being a starting/ending point for each clip and choosing the time span with the highest joint probability as the prediction [38]. In this work, we follow [23, 39] to employ a proposal-based grounding head. And in contrast to most temporal grounding works that collapse the spatial dimension, we also incorporate spatial cues to enhance temporal localization.

Text-conditioned Object Detection. Given an image, text-conditioned object detection aims to localize the target object indicated by a sentence (also named referring image expression task) or all objects described in a sentence (also named phrase grounding task). Most existing works [32,37] develop their models based on off-the-shelf detection models and match the detected objects with the sentence query. Recently, MDETR [10] and GLIP [12] unified detection and phrase grounding tasks, and they trained their models with large-scale data. These models can achieve excellent performance on text-conditioned object detection and are used to initialize model weights for STVG task [23, 30]. However, these models lack the ability of modeling dynamic cues. In this work, we design a novel collaboration framework to capture both static and dynamic cues for STVG task.

Two-stream Models. Considering that static and dynamic visual cues are quite different but complementary in video understanding, a number of two-stream models [4, 5, 21, 25, 26, 28] are proposed to capture these two kinds of visual cues. For example, Two-Stream ConvNet [21] introduces an optical flow stream to capture object motion. SlowFast [4] develops slow and fast pathways with different temporal resolutions to capture static and dynamic visual cues. However, these works typically fuse different streams equally via some simple symmetric operations, e.g., ensembling the output, sum/concat/fc/temporal-convolution on the features. In contrast, we propose a novel asymmetrical cross-stream collaboration block to exploit better collaboration between the two streams considering their different abilities.

3. Collaborative Static-Dynamic VL Streams

We illustrate our framework in Figure 2. Our framework mainly consists of a static and a dynamic VL (vision-language) stream to model static and dynamic visual-linguistic dependencies for complete cross-modal understanding. The static stream performs cross-modal understanding for static contextual information, i.e., finding the object that matches the query text in a still frame according to static visual cues, like object appearance which is important for achieving accurate spatial grounding. The dynamic stream performs cross-modal understanding for dynamic contextual information, i.e., finding the temporal moment that best matches the query sentence according to dynamic visual cues, like action which is important for achieving accurate temporal grounding. Both the static and dynamic streams learns cross-modal correspondence with a transformer-like architecture that contains multiple layers. To enable information transmission between the two streams, we further design a novel cross-stream collaboration block which is placed after each transformer layer in the two streams. In this way, both the static and dynamic VL streams can absorb complementary information from each other and achieve collaborative inferring, which can greatly reduce the uncertainty in ambiguous and hard cases where different objects have similar appearances or perform similar actions. Finally, we employ the prediction head to predict a spatio-temporal tube. In the following, we will introduce each component in detail.

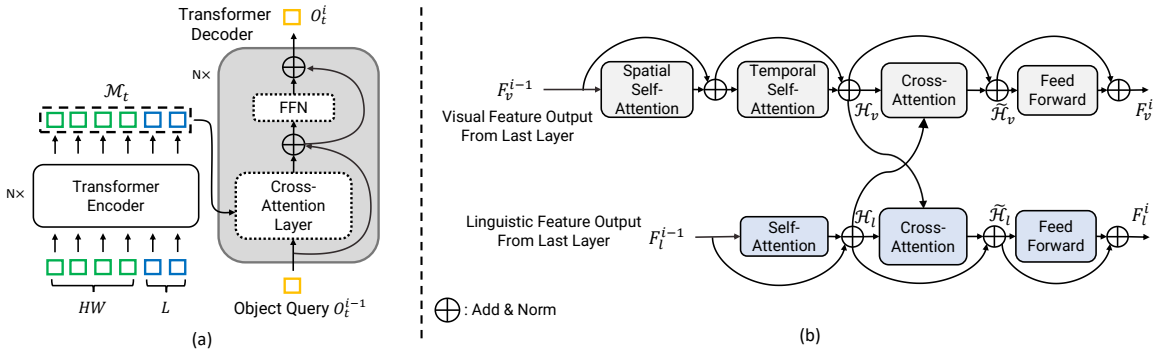

Figure 3. The detailed architectures of the proposed framework. (a): The cross-modal transformer in the static VL stream. (b): The architecture of each space-time cross-modal transformer layer in the dynamic VL stream.

3.1. Static VL Stream

The static VL stream is employed to perform cross-modal static context understanding and it processes each frame independently. Inspired by DETR [3], we design our static VL stream as a stack of N cross-modal transformer encoder layers and N cross-modal transformer decoder layers, as presented in Figure 2 and 3 (a). For the t-th frame, the inputs of our cross-modal transformer encoder are the concatenation of $\mathbb{R}^{HW\times d}$-sized visual features and $\mathbb{R}^{L\times d}$-sized language features, where H, W are the resolution of the visual features extracted from the static image frame, and L is the number of text tokens in the input text query. These visual and language features are obtained by a pre-trained image encoder [9] and language encoder [14], respectively. We employ an FC layer behind each feature extractor to project the visual and language features to a shared embedding space. The outputs of the last cross-modal transformer encoder layer form a $\mathbb{R}^{(HW+L)\times d}$-sized cross-modal memory $\mathcal{M}_t$, which captures rich interactions between intra-frame static visual cues and linguistic descriptions. Then in the decoder, a learnable object query vector $O_t^0 \in \mathbb{R}^d$ is inputted to repeatedly query object appearance and location information from the memory $\mathcal{M}_t$ via a cross-attention layer. Here, the object query gradually learns to attend to the object that matched the text query.

3.2. Dynamic VL Stream

The dynamic VL stream performs cross-modal understanding for dynamic contextual information. As illustrated in Figure 2, we first extract clip-level visual features $F_{clip} \in \mathbb{R}^{T\times H\times W\times c}$ from T uniformly sampled video clips by a pre-trained 3D-CNN [4]. Then we employ an FC layer to project the channel dimension from c to d and obtain F_v^0. F_v^0 is fed into a Space-Time Cross-Modal Transformer (STCMT) which consists of N layers to model the visual-linguistic dependencies from a dynamic perspective. The detailed architecture of each layer in STCMT is illustrated in Figure 3 (b). In each layer, we first perform intra-modality self-attention for the dynamic visual features and linguistic features. For visual features, in order to reduce computation cost, we follow TimesFormer [2] to split the spatio-temporal attention into separate attentions. Denoting the visual features after self-attention as $\mathcal{H}_v \in \mathbb{R}^{T\times H\times W\times d}$ and the linguistic features after self-attention as $\mathcal{H}_l \in \mathbb{R}^{L\times d}$. We then perform cross-attention between $\mathcal{H}_v$ and $\mathcal{H}_l$ as:

$$
\begin{aligned}
Q_v^{(h,w)} = W_v^q \mathcal{H}_v^{(h,w)}, K_v = W_v^k \overline{\mathcal{H}}_v, V_v = W_v^v \overline{\mathcal{H}}_v, \\
Q_l = W_l^v \mathcal{H}_l, K_l = W_l^k \mathcal{H}_l, V_l = W_l^v \mathcal{H}_l, \\
\widetilde{\mathcal{H}}_v^{(h,w)} = \mathcal{H}_v^{(h,w)} + \text{Attention}(Q_v^{(h,w)}, K_l, V_l), \\
\widetilde{\mathcal{H}}_l = \mathcal{H}_l + \text{Attention}(Q_l, K_v, V_v), \\
\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V,
\end{aligned}
\tag{1}
$$

where $W_v^q, W_v^k, W_v^v, W_l^q, W_l^k, W_l^v$ are learnable weighted matrices for computing queries, keys and values for the attention mechanism. $\overline{\mathcal{H}}_v \in \mathbb{R}^{T\times d}$ is obtained by conducting mean pooling on $\mathcal{H}_v$ along the spatial dimensions. d_k is the dimension of the queries and keys. $\mathcal{H}_v^{(h,w)} \in \mathbb{R}^{T\times d}$ indicates the visual feature in $\mathcal{H}_v$ at spatial position (h, w). $\widetilde{\mathcal{H}}_v, \widetilde{\mathcal{H}}_l$ are the output visual and linguistic features after cross-attention, respectively. The cross-attention operation is designed to explore rich interactions between the visual and linguistic features, which enables the model to learn a powerful cross-modal representation of the depicted dynamic cues. This is essential for grounding temporal moments. After computing the cross-attention between the visual features and linguistic features, we finally employ Feed-Forward Network (FFN) to process both features. In our dynamic VL stream, rich context is learned from the two modalities, and the visual dynamic features and the linguistic features are fused well.

3.3. Cross-Stream Collaboration

To enable the static and the dynamic VL streams to learn complementary information from each other and to perform

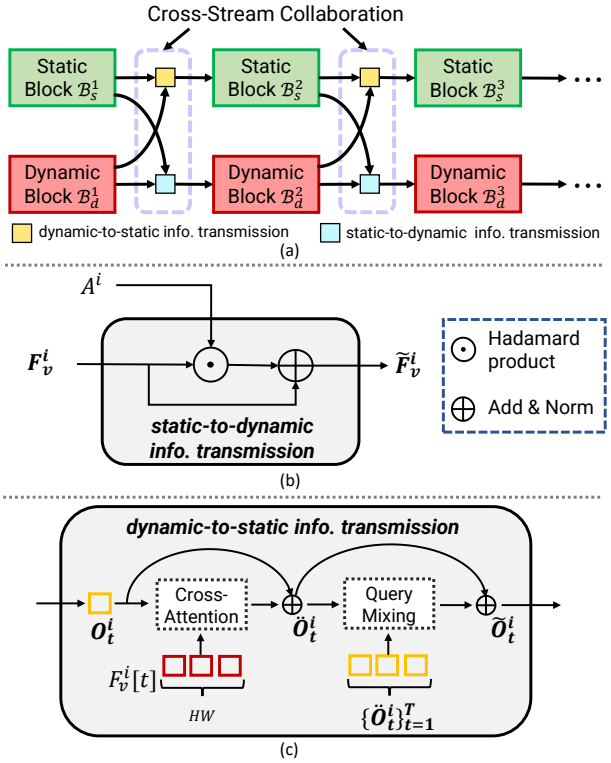

Figure 4. The architectures of our cross-stream collaboration.

collaborative reasoning, we design a novel cross-stream collaboration block between the two streams. As shown in Figure 4, this block is placed after each decoder layer $\{\mathcal{B}_s^i\}_{i=1}^N$ in static stream and each layer $\{\mathcal{B}_d^i\}_{i=1}^N$ in dynamic stream. It consists of a static-to-dynamic and a dynamic-to-static information transmission block. They are designed asymmetrically to exploit better collaboration between the two streams considering their different abilities. In the following, we will introduce the detailed designs.

The static-to-dynamic information transmission block is designed to guide the dynamic VL stream to attend to the spatial region that is highly related to the objects depicted in the query text, by utilizing the cross-attention weights learned in the decoder layers of the static VL stream. Concretely, it is formulated as follows:

$$\widetilde{F}_v^i = \text{LayerNorm}(F_v^i + A^i \odot \text{FC}(F_v^i)), \tag{2}$$

where $F_v^i \in \mathbb{R}^{T\times HW\times d}$ is the output visual feature of the i-th layer in dynamic VL stream and $A^i \in \mathbb{R}^{T\times HW\times d}$ is the cross-attention weights (replicate d times to have d channels at the last dimension) calculated between the object queries $\{O_t^i\}_{t=1}^T$ and corresponding encoded memories $\{M_t\}_{t=1}^T$ in $\mathcal{B}_s^i$. $\odot$ indicates hadamard product. The intuition behind this design is that the cross-attention weights learned in the static VL stream can attend to regions matching the language description, and this can serve as a strong guidance

Figure 5. Visualization of the learned attention map.

to help the dynamic VL stream focus more on the dynamic variations around the object-related regions. We visualize the attention maps in Figure 5. In the first sample, there are multiple boys and the attention map has a high score at the best-matched person (i.e., the tallest boy in each frame). In the second sample, we observe that when the man in red stripe is absent, the attention weights are relatively smooth. However, the attention weights become sharp once the target man appeared. These samples intuitively show that our static-to-dynamic information transmission block can guide the dynamic stream to focus on motions of the target object.

The dynamic-to-static information transmission block is designed to transfer the learned motion-text correspondence to the static VL stream. It enhances O_t^i to $\widetilde{O}_t^i$ with a cross-attention and a query mixing operation as depicted in Figure 4(c). In each block, the object query O_t^i in the static stream first queries some dynamic information from $F_v^i[t]$ (the t-th feature in F_v^i) via a cross-attention module, this operation enhances the object query representations with dynamic information learned in the dynamic stream. Then we employ a query mixing operation (implemented as temporal self-attention) to query and mix information from object queries of other frames, which further injects cross-frame information to object queries. With the proposed cross-stream collaboration block, both the static and dynamic VL streams can learn complementary information from each other in an effective way. And the enhanced representations $\widetilde{F}_v^i, \widetilde{O}_t^i$ are inputted to the next transformer layer in the static and dynamic VL streams, respectively.

3.4. Prediction Heads

In this section, we introduce our prediction heads which include a spatial and a temporal prediction head.

Spatial Prediction Head. The spatial prediction head predicts the location of the target object. For the t-th frame, the input is the object query representation $\widetilde{O}_t^N \in \mathbb{R}^d$ outputted by the last collaboration block. And we implement a 3-layer MLP to predict the bounding box location (represented by center coordinates and size) $\hat{b}_t \in \mathbb{R}^4$ of the target object.

Temporal Prediction Head. The temporal prediction head predicts the time span of the target temporal moment. Specifically, we first perform spatial mean pooling on $\widetilde{F}_v^N$

Table 1. Comparision results with previous works on HCSTVG-v1 test set and HCSTVG-v2 val set [24].

Dataset	Method	m_vIoU	vIoU@0.3	vIoU@0.5
HCSTVG-v1	STGVT [24]	18.2	26.8	9.5
	STVGBert [22]	20.4	29.4	11.3
	TubeDETR [30]	32.4	49.8	23.5
	Ours	**36.9**	**62.2**	**34.8**
HCSTVG-v2	Yu *et al.* [33]	30.0	-	-
	MMN [27]	30.3	49.0	25.6
	Aug. 2D-TAN [23]	30.4	50.4	18.8
	TubeDETR [30]	36.4	58.8	30.6
	Ours	**38.7**	**65.5**	**33.8**

Table 2. Comparison results with state-of-the-art approaches on VidSTG [42] test set.

Method	Declarative Sentences			Interrogative Sentences		
	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
STGRN [42]	19.8	25.8	14.6	18.3	21.1	12.8
STGVT [24]	21.6	29.8	18.9	—	—	—
OMRN [41]	23.1	32.6	16.4	20.6	28.4	14.1
STVGBert [22]	24.0	30.9	18.4	22.5	26.0	16.0
TubeDETR [30]	30.4	42.5	28.2	25.7	35.7	23.2
Ours	**33.7**	**47.2**	**32.8**	**28.5**	**39.9**	**26.2**

to obtain a temporal feature $F_d \in \mathbb{R}^{T\times d}$, and we use an FC layer to adjust its dimension to d_m. Then we follow 2D-TAN [39] to construct a 2D-proposal map $M \in \mathbb{R}^{T\times T\times d_m}$:

$$M_{ij} = \begin{cases} \text{MP}\left([F_d[i], F_d[i+1], ..., F_d[j]]\right) & i \le j \\ \mathbf{0} & i > j \end{cases} \quad (3)$$

where MP$(\cdot)$ is mean pooling and M_{ij} indicates the feature of temporal moment proposal C_{ij} with t_i, t_{j+1} as start and end time stamps, respectively. Here $t_i = \frac{i}{T} \cdot T_{video}$ and T_{video} is the duration of the input video. $F_d[i]$ is the i-th feature in F_d. We employ several convolutional layers to transform the 2D map M into a score map $\hat{S} \in \mathbb{R}^{T\times T\times 1}$, where $\hat{S}_{ij}$ represents the matching score of temporal proposal C_{ij}. For inference, we take the proposal with the highest score as the prediction of the target temporal time span. And we take the predicted bounding boxes from the spatial prediction head for frames within the predicted temporal span to form the final predicted tube.

3.5. Model Training

We train our model with loss $L = L_s + L_t$, where L_s and L_t are spatial and temporal localization losses, respectively. Specifically, they are defined as follows:

$$L_s = \lambda_1 L_{l1}(\hat{b}, b) + \lambda_2 L_{GIoU}(\hat{b}, b), \quad (4)$$

$$L_t = \lambda_3 L_{tg}(\hat{S}, S) + \lambda_4 L_{ta}, \quad (5)$$

where L_{l1} and L_{GIoU} are L1 loss and GIoU loss [19] on the predicted bounding boxes, respectively. The temporal grounding loss L_{tg} is defined as a binary cross-entropy loss (as done in [23, 39]) between the predicted score map $\hat{S}$ and the ground truth map S, where each element in S represents the IoU between the corresponding proposal and the ground truth temporal moment. L_{ta} is a commonly used temporal attentive loss [34] for accelerating convergence of L_{tg}, and it encourages the model to predict a high matching score for those frames/clips inside the target temporal span.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our method on HCSTVG-v1 dataset, HCSTVG-v2 dataset [24] and VidSTG [42] dataset. HCSTVG datasets are collected from movie scenes and the duration of each video clip is around 20 seconds. This set is quite challenging for spatio-temporal grounding as some video clips contain many persons conducting similar actions. HCSTVG-v1 dataset consists of 4500 and 1160 video-text pairs for training and testing, respectively. HCSTVG-v2 dataset expanded HCSTVG-v1 dataset and improved the annotation quality. It contains 10131, 2000, 4413 samples for training, validation and testing, respectively. Since the annotations of the test set are not publicly available in HCSTVG-v2, we report results on the validation set. VidSTG dataset [42] is constructed based on VidOR [20] dataset according to the object relation annotations. It contains 99943 video-text pairs, including 44808 declarative sentence queries and 55135 interrogative sentence queries. The training, validation and test set consist of 80684, 8956 and 10303 sentences respectively, and 5436, 602 and 732 videos respectively. However, since VidSTG [42] is annotated based on VidOR [20] dataset, the text queries are limited to describe the pre-defined object/relation categories in VidOR [20].

Evaluation Metrics. We follow previous works [22, 23, 30, 42] to use mean vIoU as the main evaluation metric. vIoU is defined as $\frac{1}{|T_u|}\sum_{t\in T_i} IoU(\hat{b}_t, b_t)$, where T_i and T_u indicate the intersection and union between the time intervals obtained from ground truth annotation and model prediction, respectively. $\hat{b}_t, b_t$ are the predicted bounding box and ground truth bounding box for the t-th frame, respectively. We average the vIoU score over all the samples to obtain mean vIoU (termed m_vIoU). We also report vIoU@R which indicates the proportion of samples with vIoU higher than R. In ablation studies, we follow TubeDETR [30] to also report the sIoU and tIoU metrics that evaluate the spatial and temporal grounding performances, respectively.

They are defined as $sIoU = \frac{1}{|T_{gt}|}\sum_{t\in T_i} IoU(\hat{b}_t, b_t)$ and $tIoU = \frac{|T_i|}{|T_u|}$, where T_{gt} indicates the set of frames inside the ground truth time interval.

Implementation Details. We use ResNet101 [9] as our image encoder and Roberta-base [14] as the text encoder to extract image visual features and language features. And we use the Slowfast [4] model pre-trained on AVA [8] as the video encoder. We initialize part of the weights of our static VL stream using pre-trained MDETR [10] as done in [23, 30]. The number of transformer layers in both streams is set as $N = 6$. The loss weights are set as $\lambda_1 = 5, \lambda_2 = 2, \lambda_3 = 5, \lambda_4 = 1$. We train our model on VidSTG [42], HCSTVG-v1, HCSTVG-v2 [24] datasets for 10, 10, 4 epochs, respectively. For most other hyper-parameters, we keep them consistent with previous works [10, 30]. Further implementation details can be found in the supplementary material.

4.2. Experimental Results

In this section, we compare the performance of our approach with previous methods on HCSTVG dataset [24] and VidSTG [42] dataset.

HCSTVG datasets. We first compare our method with state-of-the-arts on HCSTVG datasets [24] and results are shown in Table 1. On HCSTVG-v1 test set, our approach outperforms previous state-of-the-art method TubeDETR[1] [30], by a considerable margin. We achieve 36.9% m_vIoU, which is 4.5% higher than TubeDETR. On HCSTVG-v2 validation set [24], we compare our method with TubeDETR [30] and some winners from HCVG challenge2021. As shown in Table 1, we also outperform TubeDETR by 2.3% m_vIoU. The above results demonstrate that our proposed model with two collaborative vision-language streams can effectively capture static-dynamic cross-modal dependencies for addressing STVG task.

VidSTG dataset. We also report results on VidSTG dataset [42]. As listed in Table 2, we also achieve state-of-the-art performance with a m_vIoU of 33.7% on declarative sentences and 28.5% on interrogative sentences, respectively. This further shows the effectiveness of the proposed collaborative two-stream model.

4.3. Ablation Study

In this section, we conduct ablation experiments on HCSTVG-v2 dataset [24] to verify the effectiveness of some important components in the proposed method.

Quantitative evaluation on the Cross-Stream Collaboration Block. We first verify the effectiveness of static-to-dynamic and dynamic-to-static information transmission blocks by removing one of them from our developed framework. We present the evaluation results in Table 3. Compared to the baseline in which the whole cross-stream collaboration block is removed (the first line in Table 3), it is clear that individually adopting static-to-dynamic information transmission (termed "s-to-d") or dynamic-to-static information transmission (termed "d-to-s") can both improve the grounding performance on vIoU metrics. We also observe that "s-to-d" mainly improves the temporal grounding accuracy while "d-to-s" improves the spatial grounding performance. And our full model that jointly employs the whole cross-stream collaboration block achieves the best performance of 38.7% m_vIoU. The above results show that both the static-to-dynamic and the dynamic-to-static information transmission blocks are effective for providing complementary information to the other stream, and thus employing them greatly improves the performance.

Table 3. Evaluation on the cross-stream collaboration mechanism.

s-to-d	d-to-s	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_sIoU
		36.4	61.2	29.5	56.1	64.9
✓		37.5	63.6	31.4	57.4	64.6
	✓	37.1	62.3	30.2	56.2	65.7
✓	✓	**38.7**	**65.5**	**33.8**	**58.1**	**65.7**

Qualitative evaluation on the Cross-Stream Collaboration Block. To better understand how and why our cross-stream collaboration block significantly improves the grounding accuracy, we further provide 2 qualitative samples in Figure 6. In the first sample, there are three boys coming to the dining table. In this case, our baseline model without cross-stream collaboration failed to predict correct temporal grounding results since it is confused by the similar action performed by the three boys. With our proposed cross-stream collaboration block, the dynamic VL stream can distinguish which boy is the "tall boy" by absorbing knowledge from the static VL stream, then it can correctly predict the target temporal time span. For the second sample, there are two women both dressed in a skirt, thus in our baseline where the cross-stream collaboration is removed, the static stream failed to capture and understand the action "let goes of ... and turns into the room", thus it cannot distinguish which is the target woman. By employing the proposed cross-stream collaboration block, the motion information learned in the dynamic VL stream can be transmitted to the static stream to make it aware of the motion information which can help distinguish the target person. The above two samples clearly show that the proposed cross-stream collaboration block can effectively exchange complementary information between the static and the dynamic VL streams, which can greatly reduce the uncertainty in the hard and ambiguous cases where some objects have a similar appearance or motion so that the model can produce a more accurate prediction.

Spatial information in the dynamic VL stream. We

[1] In this paper, we report the corrected TubeDETR [30] performance updated by TubeDETR authors at https://github.com/antoyang/TubeDETR.

Figure 6. Visualization of the predicted tube of our approach (yellow), our approach without cross-stream collaboration (red), and ground truth (green). In the first sample, our approach without cross-stream collaboration predicts a wrong temporal time span. In the second sample, ours w/o the collaboration block predicts wrong spatial bounding boxes. (Best viewed zoomed in on screen.)

Table 4. Evaluation on designs of the dynamic VL stream.

Method	Ours	Ours w/o spatial	Ours w/o mean pool
m_tIoU	**56.1**	54.7	**56.2**

Table 5. Comparision between different cross-stream collaboration block designs.

Method	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_sIoU
Ours	**38.7**	**65.5**	**33.8**	**58.1**	**65.7**
None	36.4	61.2	29.5	56.1	64.9
sum	36.6	62.5	30.0	55.6	64.4
FC + sum	36.8	62.8	30.5	57.3	63.4
concat + FC	37.1	63.2	30.3	56.9	63.9

present ablation studies on different design choices of the dynamic VL stream and results are presented in Table 4. Here, we discard the static stream to eliminate its influence and we report temporal grounding accuracy using m_tIoU metric. By mean pooling at the spatial dimension of the video feature inputted to the dynamic VL stream (termed "Ours w/o spatial"), it results in a drop of more than 1% m_tIoU which shows the importance of learning spatial visual interaction for modeling dynamic cues. And since we develop the dynamic VL stream to mainly learn dynamic cross-modal correspondence, we employ spatial mean pooling on $\mathcal{H}_v$ to obtain $\overline{\mathcal{H}}_v$ so that the text-to-visual cross-attention is performed only on the temporal dimension which can greatly reduce computation cost. By removing this mean pooling operation (termed "Ours w/o mean pool"), the text-to-visual cross-attention is performed among all spatio-temporal tokens, but it has very limited improvement on the performance, which validates the rationality of our design.

Further study of the collaboration block. To further verify the effectiveness of the proposed collaboration block, we replace it with some symmetric operations, including "sum", "FC + sum" and "concat + FC", which are some operations commonly used in the community [4, 21]. As shown in Table 5, "concat + FC" achieves the best performance among these baselines, but it only achieves a limited improvement of 0.7% m_vIoU, which indicates that simply merging the information from the two streams can not fully exploit their complementary abilities. In contrast, "Ours" outperforms "concat + FC" by a considerable margin of more than 3% in terms of vIoU@0.5, demonstrating that our asymmetric cross-stream collaboration block enables the two streams to effectively exchange complementary information and to collaboratively reason the target object.

5. Conclusion

In this work, we propose a novel framework for Spatio-Temporal Video Grounding task. Our framework mainly consists of a static VL(vision-language) stream which captures static visual cues like object appearance, and a dynamic VL stream which understands dynamic visual cues like action. We elaborately design the framework as a collaborative system in which the two streams collaborative localize the target object via a novel cross-stream collaboration block, which is proven to be effective. We also present visualization results which provide some interesting insights and show how the collaboration of the two streams helps improve the prediction. Our approach significantly outperforms previous methods on VidSTG [42] and HC-STVG dataset [24], which demonstrates its effectiveness.

Acknowledgements. This work was supported partially by the NSFC (U21A20471, U22A2095, 62076260, 61772570), Guangdong Natural Science Funds Project (2020B1515120085, 2023B1515040025), Guangdong NSF for Distinguished Young Scholar (2022B1515020009, 2018B030306025), and the Key-Area Research and Development Program of Guangzhou (202007030004). We thank Bing Shuai for the helpful discussions.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 2, 3, 4, 7, 8

[5] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 3

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[8] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7

[10] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3, 7

[11] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1902–1910, 2021. 2, 3

[12] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[13] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 1

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 7

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[16] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2, 3

[17] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775, 2021. 2

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[20] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 6

[21] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1, 2, 3, 8

[22] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1533–1542, October 2021. 1, 2, 6

[23] Chaolei Tan, Zihang Lin, Jian-Fang Hu, Xiang Li, and Wei-Shi Zheng. Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. *arXiv preprint arXiv:2106.10634*, 2021. 3, 6, 7

[24] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021. 1, 2, 6, 7, 8

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 3

[26] Xuanhan Wang, Lianli Gao, Peng Wang, Xiaoshuai Sun, and Xianglong Liu. Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Transactions on Multimedia*, 20(3):634–644, 2017. 3

[27] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. *CoRR*, abs/2109.04872, 2021. 6

[28] Zuxuan Wu, Yu-Gang Jiang, Xi Wang, Hao Ye, and Xiangyang Xue. Multi-stream multi-class fusion of deep networks for video classification. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 791–800, 2016. 3

[29] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021. 2

[30] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 6, 7

[31] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 1

[32] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 1, 3

[33] Yi Yu, Xinying Wang, Wei Hu, Xun Luo, and Cheng Li. 2rd place solutions in the hc-stvg track of person in context challenge 2021. *arXiv preprint arXiv:2106.07166*, 2021. 6

[34] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019. 6

[35] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2, 3

[36] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry 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*, pages 1247–1257, 2019. 2

[37] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 3

[38] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2, 3

[39] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020. 2, 3, 6

[40] Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Chuanqi Tan. Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):3130–3139, 2019. 1

[41] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. 2, 6

[42] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6, 7, 8