Human-centric Spatio-Temporal Video Grounding With Visual Transformers

Zongheng Tang, Yue Liao, Si Liu*, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, Dong Xu, *Fellow, IEEE*

Abstract—In this work, we introduce a novel task – Human-centric Spatio-Temporal Video Grounding (HC-STVG). Unlike the existing referring expression tasks in images or videos, by focusing on humans, HC-STVG aims to localize a spatio-temporal tube of the target person from an untrimmed video based on a given textural description. This task is useful, especially for healthcare and security related applications, where the surveillance videos can be extremely long but only a specific person during a specific period is concerned. HC-STVG is a video grounding task that requires both spatial (where) and temporal (when) localization. Unfortunately, the existing grounding methods cannot handle this task well. We tackle this task by proposing an effective baseline method named Spatio-Temporal Grounding with Visual Transformers (STGVT), which utilizes Visual Transformers to extract cross-modal representations for video-sentence matching and temporal localization. To facilitate this task, we also contribute an HC-STVG dataset[1] consisting of 5,660 video-sentence pairs on complex multi-person scenes. Specifically, each video lasts for 20 seconds, pairing with a natural query sentence with an average of 17.25 words. Extensive experiments are conducted on this dataset, demonstrating that the newly-proposed method outperforms the existing baseline methods.

Index Terms—Spatio-Temporal grounding, transformer, dataset

I. INTRODUCTION

GIVEN natural language queries, a visual grounding task aims to localize objects or regions in images or videos, which is an important task in the vision-language research field. It was originated by localizing objects in an image based on short descriptions or sentences [1], [2] (Fig. 1(a)). With the progress on video understanding, recent efforts have been made on referring expression in videos, such as temporal video grounding. Taking Fig. 1(b) for an example, this video grounding task aims to localize a video segment corresponding to the given language query.

Several existing tasks are relevant to the newly proposed HC-STVG task from the task perspective, but they have a different focus. Specifically, **Referring Expression** aims to spatially localize the target object in an image and outputs a bounding box. **Temporal video grounding** aims to determine

the temporal boundaries of the target event in a video and thus returns a set of consecutive frames. Compared with these two tasks, the HC-STVG task aims to localize the spatio-temporal tube of the target person described in the textual query, which is more challenging as it requires localization both spatially and temporally. For example, the target person may appear in the video for a long time, but the interval of the event matching the description may be short. It is necessary to utilize both the textual cues (such as actions and relations with other objects) indicated in the query sentence and the visual cues shown in the video to determine the temporal boundaries. Furthermore, compared to the **STVG** task, we only consider humans as our referent. This is out of two considerations. First, humans are specific and always the focus of all species in the videos. Second, humans produce a wealth of action information and interact with other people or things. Therefore, the HC-STVG task requires a model to do fine-grained reasoning, which is more challenging but have more practical applications in the real world.

Grounding in both space and time domain is often required in real-world scenarios, especially for healthcare and security related applications, where the surveillance videos can be extremely long, but people only care about a specific person and his/her specific behaviors. For example, in a 24-hours monitoring video, a healthcare doctor at the nursing home may want to see how a particular older person was having his/her lunch, or a police officer may need to search for a suspect from the crowded people and locate the process when the suspect conducts a crime. In both cases, when the description is provided, the expected outputs are a sequence of bounding boxes related to the target person in the consecutive frames corresponding to the described action, i.e., a spatio-temporal tube. Unfortunately, such a video grounding setting has not been explored yet. Hence the existing methods cannot well handle this task.

In this work, we introduce a new task named Human-centric Spatio-Temporal Video Grounding (HC-STVG), which aims to localize a spatio-temporal tube of the target person in an untrimmed video given a query description. As shown in Fig. 1(c), based on the description 'The woman in a white apron comes to the table and puts the tray in her hands on the table', an HC-STVG model needs to answer: which person is the target and which frames correspond to the described actions. To the best of our knowledge, it is the first video grounding task centered on humans. There are three key challenges in our HC-STVG task: 1) It requires grounding

Zongheng Tang, Yue Liao, Si Liu, Qian Yu, and Hongxu Jiang are with Beihang University, Beijing, China. Guanbin Li is with the School of Computer Science and Engineering, Sun Yat-sen University, China, and Pazhou Lab, Guangzhou, 510330, China. Xiaojie Jin is with ByteDance AI Lab, Beijing, China. Dong Xu is with the School of Electrical and Information Engineering, the University of Sydney, Sydney, Australia. The Corresponding author is Si Liu (Email: liusi@buaa.edu.cn).

[1]The new dataset is available at https://github.com/tzhhhh123/HC-STVG.

(a) Referring expression in images[1,2]

(b) Temporal language grounding [3,4]

(c) Human centric **spatio-temporal** referring (ours)

Fig. 1. Comparison of different visual grounding tasks. (a) Referring expression in images (e.g., [1], [2]) is a task of localizing objects/regions in images based on a given query. (b) Temporal video grounding aims at localizing the starting and ending frame corresponding to the described action(s) (e.g., [3], [4]). (c) Human-centric Spatio-temporal Video Grounding is our newly proposed task, which outputs both spatial and temporal localization. The example is from our newly collected HC-STVG dataset. Each example includes a video clip and a corresponding description (top), spatial annotation (i.e., the green boxes), and temporal annotations (i.e., the blue bottom line).

referring expressions both spatially and temporally at the same time. 2) Multi-modality information, such as visual/textural attributes and actions, is often required for localizing a specific person, especially in a complex multi-person scene. 3) It can be difficult to determine the starting and ending frame of an action due to its dynamic nature. This task is meaningful as it involves cross-modality modeling and fine-grained reasoning.

To effectively deal with this task, we propose a new baseline method, termed Spatio-Temporal Grounding with Visual Transformers (STGVT), in which we jointly exploit multi-modality information from videos and textural descriptions by using a visual Transformer. Specifically, our STGVT method consists of four steps: After generating tube proposals from a video by linking bounding boxes in consecutive frames, our method learns the cross-modality interaction between the tube proposals and the query sentence via a Visual Transformer, which is then followed by predicting the matching tube and trimming the irrelevant frames. Although there are many vision-language datasets for the visual grounding tasks, none can support this new task. To address the issue, we contribute an HC-STVG dataset. Precisely, it consists of 5,660 video-query pairs, where 57.2% of scenes have more than three people. Each query is a sentence with an average of 17.25 words, consisting of rich expressions related to actions and human-object interaction. Based on this new benchmark dataset, we demonstrate the effectiveness of our baseline method for the newly proposed HC-STVG task.

To summary up, the contributions of this work are three-fold: 1) We introduce a novel and challenging task,

i.e., Human-centric Spatio-Temporal Video Grounding (HC-STVG), which for the first time, focuses on humans in spatio-temporal video grounding. 2) We build the first human-centric video-description dataset, which can be used as the benchmark dataset for the new task and inspires the subsequent research. 3) We propose an effective baseline method for the new task and demonstrate the effectiveness of the proposed method through extensive experiments.

II. RELATED WORK

A. Referring Expression in Images/Videos

Referring Expression aims to localize the visual object described by any natural language expression [1], [2], [6]–[12]. The previous works [6] and [11] used a pretrained object detection network or an unsupervised method to generate object proposals and then match these regions with the textual description to select the most relevant proposal. To better capture multi-modality context information, MAttNet [11] proposed to decompose the referring expression into *subject*, *location*, and *relation*. Moreover, cross-modality correlation filtering is adopted for the referring expression task in real-time [12]. As for referring expression in videos, the WSSTG [13] is proposed to localize the spatial-temporal tube of an object in a *trimmed* video where the target objects exist through the whole clip.

B. Visual Transformer

In recent years, Visual Transformer methods [14]–[20] have achieved impressive performance in many Vision-Language

The woman in black clothes comes to a table, stops, and puts her hand on the table.

14.04s ⟷ 17.55s

(a) An example of our newly collected HC-STVG dataset

The man in the vest puts the plate on the table and turns away

| Annotate | Track | Annotate | Track | Annotate |

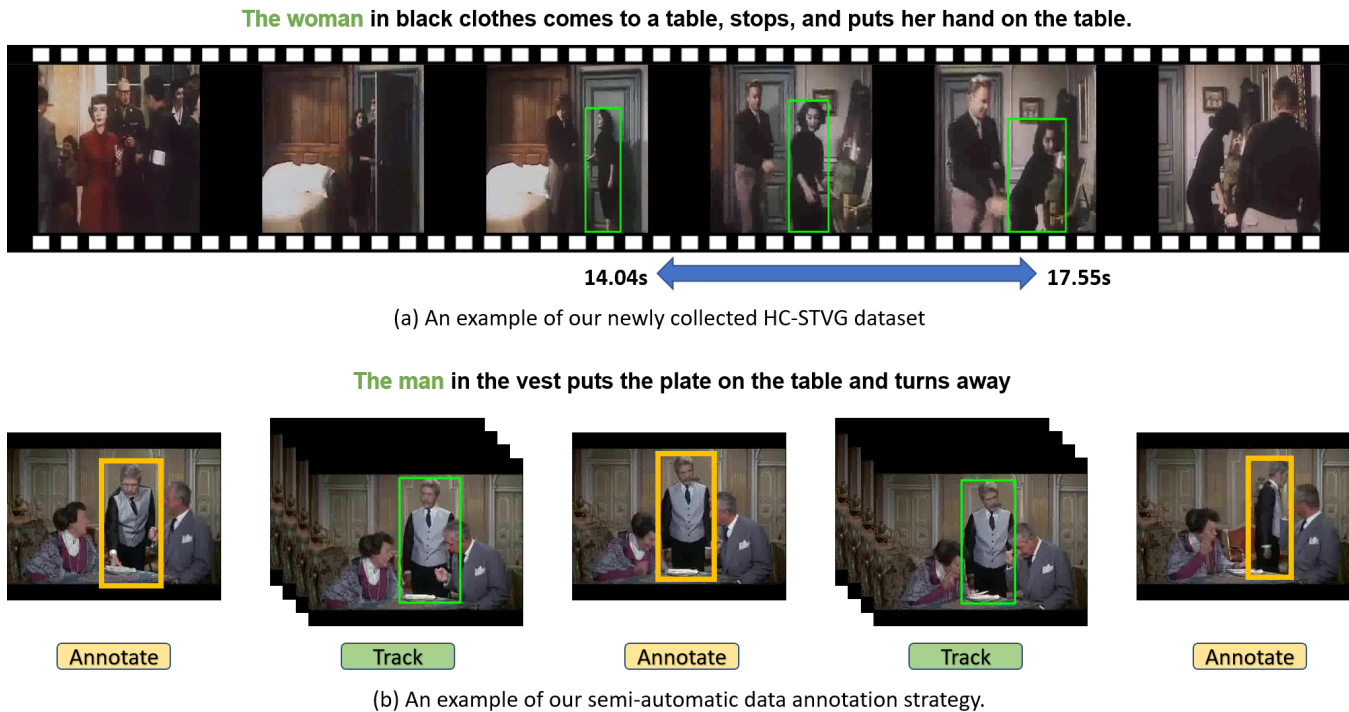(b) An example of our semi-automatic data annotation strategy.

Fig. 2. Examples of our newly collected HC-STVG dataset. (a) For a video clip, we collect its text description (top), bounding boxes (green boxes), and temporal annotation (bottom). (b) An example of our semi-automatic annotation strategy. For each video clip, the volunteers are asked to annotate the yellow bounding boxes in three selected keyframes, while the green bounding boxes in the remaining frames are tracked by the method SiamPRN [5].

tasks, including the Referring Expression tasks. These works usually employed a transformer-structure model to extract multi-modal feature representation from image-text or video-text pairs and then predict the regions of interest(ROI) based on the cross-modal features. The existing transformer-based methods proposed for vision-language tasks can be divided into two categories, the single-stream models and the two-stream models. The single-stream models [14]–[18] employ a single transformer to learn cross-modal features from the visual and the textual modalities. In contrast, the two-stream models [19], [20] use different encoders to encode information of different modalities while using a co-attention module to learn cross-modal feature representation.

C. Temporal Video Grounding

The goal of temporal video grounding is localizing the most relevant video segment given a query sentence. The previous works [3], [4], [21]–[23] used a temporal sliding-window-based approach over the video frames to generate temporal candidates and choose the most relevant one as the output. Chen et al. [24] proposed aggregating the fine-grained frame-by-word interaction between videos and queries. Xu et al. [23] and Chen et al. [25] proposed to generate query-specific proposals as candidate segments by directly integrating sentence information with each fine-grained video clip. Ge et al. [26] explored activity concepts in both videos and queries for temporal localization. Zhang et al. [27] iteratively adjusted the structured graph to deal with the semantic misalignment problem. Yuan et al. [28] utilized the Graph Convolutional Network [29] to model the relations between different video

clips and proposed a temporal conditioned pointer network to screen the answer. Unlike these previous video grounding works, which only deal with temporal localization, the newly proposed HC-STVG tackles spatial localization and temporal localization simultaneously.

III. HC-STVG BENCHMARK

The newly proposed HC-STVG task aims to localize the target person spatio-temporally in an untrimmed video. For this task, we collect a new benchmark dataset with spatio-temporal annotations related to the target persons in complex multi-person scenes, together with full interaction and rich action information.

A. Task Formulation

In this work, the HC-STVG task is defined as follows. Given an untrimmed video $v \in V$ and a natural language description $s \in S$, which describes a sequence of actions related to a specific person in v, we aim to generate the spatio-temporal tube T (i.e., a sequence of bounding boxes) corresponding to the target person. Specifically, the system needs to locate the segment by determining the starting and ending frames (l, r) and also generates the bounding boxes of the target person in the located segment.

B. Overview of HC-STVG dataset

There are 5,660 video-sentence pairs in our newly collected HC-STVG dataset, including 4,500 training pairs and 1,160 testing pairs. The average temporal duration of the ground-truth tubes is 5.37 seconds while each sentence has an average

of 17.25 words. The duration of videos is further normalized to 20 seconds. We ensure that the test samples and the training samples do not originate from the same raw video.

The key characteristics of our HC-STVG are summarized as follows: (1) human-centric. The dataset contains precise spatio-temporal annotations and textual description for the person of interest in each video. (2) All videos are captured in complex multi-person scenes, along which 57.2% of the videos have more than three people while the rest videos have two persons. (3) All sentences have rich descriptions related to the interaction between human-human or human-objects and 56.1% of the descriptions include both types of interaction.

Figure 2(a) and Fig. 3 show several examples of our newly collected dataset. As demonstrated in these two figures, the videos in the new dataset are captured in multi-person scenes. Besides, the descriptions have rich expressions related to actions and relations.

C. Dataset Construction

The whole process includes five modules: Raw Video Preparation, Video Span Selection, Video Description, Bounding box annotation, and Video Span Extension. We explain these parts in detail.

Raw Video Preparation We use raw video filtered by AVA dataset [30]. These original videos have been collected on Youtube and have been carefully selected to ensure the diversity and high-quality of the videos. These videos contain a lot of realistic scenes, character actions, and character interactions. Therefore, we can produce more suitable video-description pairs from the videos.

Video Span Selection Annotators watch untrimmed videos sequentially and follow several rules to mark suitable describable spans. To be specific, the chosen spans must be multi-person scenes, thus increasing the difficulty of grounding in our dataset. Shot switching is not allowed in the annotated clips because shot switching splits the video content into discrete paragraphs, bringing semantic incoherence. And We limit the shortest video span time to 3s to avoid annotators only tagging the instantaneous actions of the characters instead of continuous action sequences. Under these restrictions, annotators annotate the precise temporal segment of the target person's actions in the scenario.

Video Description The annotator who decides the specific video span is required to write down a description of the target person. To collect more diversified and natural expressions, we do not provide any templates and predefined verb lists for annotators to follow. But we still have some guidelines for annotators to make an appropriate textual annotation. (1) The description can spatio-temporally locate the target person uniquely, which means that there can not exist another person that satisfies the description and the actions of the target person are traceable. (2) We encourage annotators to describe in terms of explicit visual attributes, human actions, and relations, as our goal is to contain a variety of reasoning information. And descriptions of other persons' attributes in the sentence are also encouraged to increase diversity. (3) As for actions descriptions, in addition to using some transitive verbs (e.g. walk, turn) for target persons, workers are encouraged to write

human-to-human and human-to-object interactions (e.g. hug, follow) as much as possible, which urges the algorithm to pay close attention to the identification and alignment of dynamic visual relation. (4) The annotators are encouraged to label successive events that consist of the target character and related actions, which increases the number of verbs in the sentence and differs HC-STVG from the mere action localization task. (5) Since we only deal with the video images, the annotators are required to focus on the visual content rather than the sound content. At last, after the annotations, we manually review the descriptions to ensure label quality.

Bounding Box Annotation Frame-level bounding box annotation is quite crucial for visual grounding tasks to assist in the judgment of location correctness. However, manually labeling bounding boxes for each frame in the video may lead to enormous labor pressure. To trade off this problem, we adopt the way of manually labeling keyframes and automatically labeling other frames with the tracking algorithm. Concretely, we use SiamRPN [5] to track the target person. We manually annotate three keyframes of a video span (i.e., starting, middle, ending frame) and track the other frames, which effectively outperforms using only one keyframe. We track the person in the annotated temporal clip from front to back and from back to front using keyframes, then we average these two tracking results as annotation. We conduct an additional manual review to correct the tube for samples with apparent differences in two direction tracking results.

Video Span Extension After the video spans are collected and the spatial-temporal annotations are complete, we extend the video clips forward and backward on the temporal ground truth randomly to achieve a fixed total length of time for final query videos, and the extended time also serves as a negative sample for temporal localization. As localizing descriptions in query videos may be ambiguous after time extension, in HC-STVG, additional annotators review the total video span to guarantee there is no unambiguous referring.

TABLE I
STATISTICS OF DIFFERENT DATASETS.
ALS STANDS FOR AVERAGE LENGTH OF SENTENCES. BA AND TA STAND FOR BOUNDING BOX ANNOTATION AND TEMPORAL ANNOTATION, RESPECTIVELY. HC STANDS FOR HUMAN-CENTRIC.

Dataset	#Queries	#Videos	ALS	BA	TA	HC
DiDeMo [4]	40,543	10,464	8.0	×	✓	×
CharadesSTA [3]	16,128	6,670	7.2	×	✓	✓
TACoS [31]	18,818	7,206	10.5	×	✓	✓
ActivityNet-C [32]	71,942	12,460	14.8	×	✓	✓
VID-sentence [13]	7,654	5,318	13.2	✓	×	×
VidSTG [33]	44,808	6,924	11.12	✓	✓	×
HC-STVG Dataset	5,660	5,660	17.25	✓	✓	✓

D. Comparison with the Existing Datasets

We compare the existing video grounding datasets with our newly collected HC-STVG dataset in Table I. The DiDeMo dataset [4] only provides temporal annotation for the temporal localization task. The TACoS dataset [34], ActivityNet Captions dataset [32], and Charades-STA dataset [3] are all human-centric datasets as their videos capture human actions

Fig. 3. An example video-sentence pair in our newly collected dataset. In the example video clip, the target person performs a series of actions. Correspondingly, these actions are described in the sentence, including 'puts his glasses on the table','pulls something out of his pocket', and 'puts it on the table'.

(e.g., cooking). Similar to DiDeMo dataset, these datasets are collected for the temporal localization task, and they do not have annotations at the bounding box level. The VID-sentence dataset [13] provides the bounding-box-level annotations. However, it only focuses on spatial localization. Video clips in this dataset are all trimmed, hence it is not suitable for temporal localization. Among all datasets, the most relevant dataset is the VidSTG dataset [33]. It is extended from the dataset Vidor [35], which is a dataset originally collected for detecting relations in videos. The VidSTG dataset provides both bounding-box-level annotations and temporal annotations. However, this dataset focuses on the task of relation referring between subjects and objects. In contrast, our HC-STVG dataset focuses on localizing the target person spatially and temporally from multi-person scenes based on the query sentences. Furthermore, on average the length of each sentence in our dataset is the longest among these datasets, and the descriptions from this dataset contain rich expressions related to actions and human-object interaction. The details of all these datasets are summarized in Table I.

IV. METHODOLOGY

The overall framework of our STGVT method is shown in Fig. 4. Given a video v and a query sentence s depicting a person performing a series of actions, our STGVT method aims to output a spatio-temporal tube T corresponding to the description. The method consists of four steps: **1)** it first generates the tube proposals (Sec. IV-A). **2)** Then a visual transformer takes a pair of the tube proposal and the query sentence as the input, and extracts the cross-modal features, including the global feature and the frame features (Sec. IV-B). **3)** A matching classifier is adopted to predict a matching score for each tube-sentence pair. The tube proposal with the highest matching score will be proceeded in the final step (Sec. IV-C). **4)** Based on the selected tube, a temporal trimming module is proposed to trim the irrelevant frames and produce the final spatio-temporal localization results. (Sec. IV-D).

In this section, we first describe each step of the proposed method and then introduce the loss functions (Sec. IV-E) and the inference process (Sec. IV-F).

A. Tube Proposals Generation

Given a video, we first use the region proposal network [36] to detect the bounding boxes $B_{t,i}$ in each frame, where $B_{t,i}$ represents the i-th bounding box in the t-th frame. Basically, we follow the ACT method [37] to link the bounding boxes in consecutive frames to form spatio-temporal tube proposals T_p. This process is essential to group the bounding boxes contain-
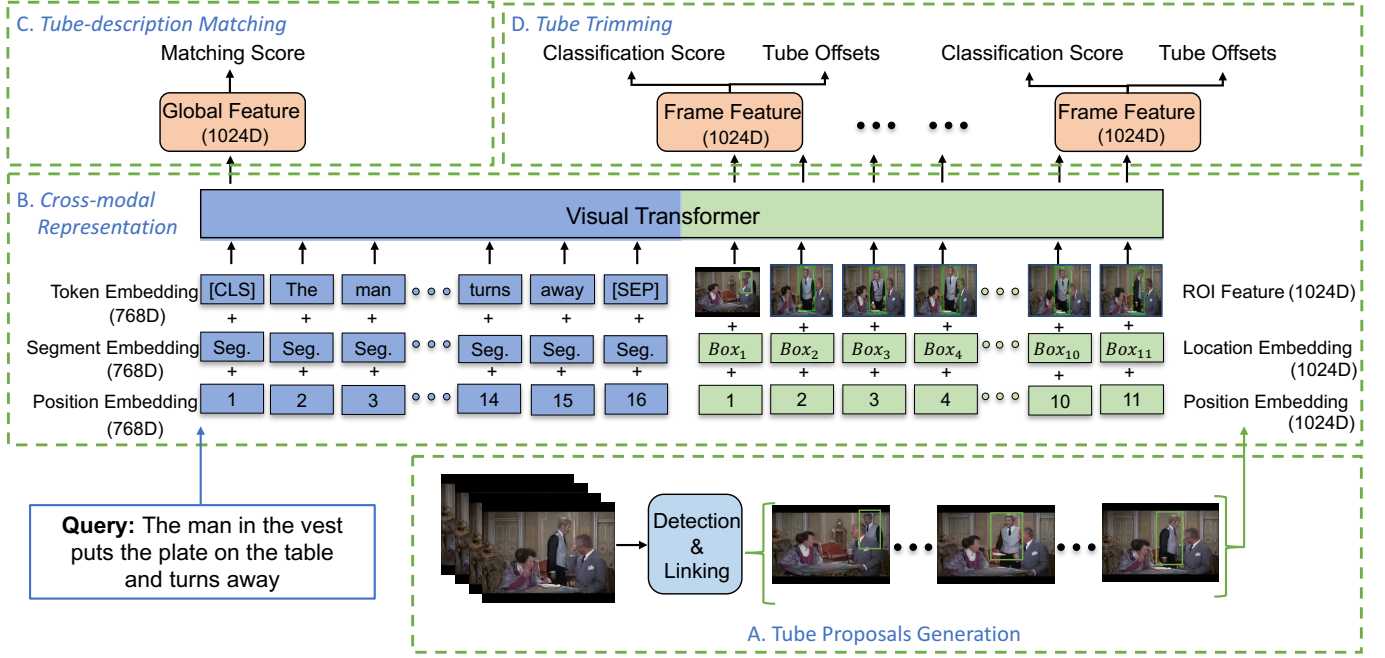
Fig. 4. The framework of our STGVT. The framework has four modules, which correspond to four steps explained in Methodology. Specifically, given a query sentence and video, it first generates tube proposals (A), and then each tube-description pair is fed into the visual transformer (B). The output feature of the visual transformer is used for predicting matching score, classification score, tube offsets , which to be used for tube selection (C) and tube trimming (D).

ing similar visual content in adjacent frames. Different from the method [37], we compute the similarity score between two bounding boxes (i.e., $B_{t,i}$ and $B_{t+1,j}$) as follow:

$$
\begin{aligned}
S(B_{t,i}, B_{t+1,j}) = & \lambda_i * IoU(B_{t,i}, B_{t+1,j}) \\
& + \lambda_c * D_{cosine}(f(B_{t,i}), f(B_{t+1,j})) \\
& + \phi(B_{t,i}) + \phi(B_{t+1,j}),
\end{aligned} \tag{1}
$$

where $f(\cdot)$ is a pretrained feature extractor, $D_{cosine}(\cdot, \cdot)$ is the cosine function, ϕ is the confidence score of the bounding boxes, and λ_i, λ_c are the weights of the IoU score and the cosine score. In practice, we set them as 0.7 and 0.3 respectively.

B. Cross-modal Representation

The next step is to extract cross-modal features for each pair of the tube proposal T_p and the query description s. We adopt Visual Transformer in this step due to its great success for the cross-modal tasks [19]. The visual transformer takes the visual and textual inputs separately and models their interaction through a set of transformer layers based on the co-attention mechanism. Specifically, following [38], we use the position information, token information, and segment embedding extracted from the query description s as the textual inputs; while each temporal position, visual feature, and spatial location from tube proposal T_p are used as the visual inputs for each tube proposal T_p. The temporal position refers to the frame index, the visual features are extracted from the bounding box in each frame of T_p, and the bounding boxes' coordinates are encoded as the location embedding.

The visual transformer will output two cross-modal features f^{global} and f^{frame}. When using the textual inputs as the Query and the visual inputs as the Key and Value, f_p^{global} (p indicates the p-th tube-sentence pair) is the element-wise product of the visual feature and textual feature, which are extracted from the first token position. When the visual inputs are used as Query and the textual inputs are used as Key and Value, the model generates the feature $f_{p,t}^{frame}$ for each bounding box in each frame of T_p (p and t refer to the t-th frame of the p-th tube). These two features are used to predict the matching scores of T_p at the tube-level and the frame-level, which to be explained below.

C. Tube-description Matching

In this step, our method will verify if the input tube proposal T_p is matching with the query description s. As mentioned above, the visual transformer extracts the global feature f_p^{global} for each pair of T_p and s. We feed f_p^{global} into a binary classifier to predict its matching score $M_p' \in [0,1]$. T_p with the highest matching score will be selected and fed into the last step.

When training the classifier, it is worth noting that the positive and negative samples can be imbalanced if we only treat the ground-truth tube, T_{GT}, as the positive sample. To tackle this problem, we relax the criteria for the positive samples by introducing two scores, $s_{overlap}$ and s_{IoU}. The tube proposals which simultaneously satisfy the following two conditions, $s_{overlap} >= 0.9$ and $s_{IoU} > 0.5$, are treated as the positive samples. Specifically, (1) $s_{overlap}$ is defined as follow:

$$s_{overlap} = \frac{|T_p \cap T_{GT}|}{|T_{GT}|}, \qquad (2)$$

$|T_{GT}|$ refers to the total number of frames in the ground-truth tube. This score reflects the ratio of the total number of the intersected frames between T_p and T_{GT} over all frames in $T_G T$. 2) The average IoU score s_{IoU} is defined as follow:

$$s_{IoU} = \frac{1}{|T_p \cap T_{GT}|} \sum_{t \in T_p \cap T_{GT}} IoU\left(B'_t, B_t\right), \qquad (3)$$

where B'_t and B_t are the detected/ground-truth bounding boxes of T_p and T_{GT} at frame t, respectively. A tube is treated as a negative sample when its s_{IoU} is lower than 0.2.

D. Tube Trimming

Given the dynamic characteristics of actions, the selected tube proposal T_p may contain redundant transition frames. Hence it is required to trim T_p to output the final predicted tube T. We compute the relevance score of each frame in T_p to the query s, and predict the regression offsets. Our trimming module consists of two subnets, a classification subnet and a boundary regression subnet, which will be detailed below.

1) Classification Subnet

The classification subset is used to predict if a video frame should be kept in the final tube – the spatio-temporal localization result T, or not. As introduced in Sec. IV-B, the transformer outputs $f_{p,t}^{frame}$ for each frame of T_p. The classifier takes $f_{p,t}^{frame}$ as the input to predict a relevance score $C'_{p,t}$ indicating how relevant the t-th frame is to the query sentence s.

2) Boundary Regression Subnet

Besides the relevance score $C'_{p,t}$, we also adopt a network to predict the temporal offsets, which reflects how far away the current frame is from the ground-truth temporal boundary. This subset's architecture is the same as the classification subset except that it has two outputs. Specifically, for a positive frame at the t-th position, if the ground-truth tube spans from the l-th frame to the r-th frame (i.e., $R_{GT} = [l, r]$), the regression target is $O_{p,t} = (\delta l, \delta r)$,

$$\delta l = \frac{t - l}{N} \qquad \delta r = \frac{r - t}{N}, \qquad (4)$$

where N is the number of frames in tube T_p. δl and δr refer to the temporal offset from the t-th frame to the left and the right boundary, respectively.

E. Loss Functions

Given the tube-level matching prediction score M'_p, the frame-level relevance prediction score $C'_{p,t}$, and boundary regression prediction result $O'_{p,t}$, and their corresponding ground-truth labels M_p, $C_{p,t}$, $O_{p,t}$, the total loss is defined as follows,

$$L = \sum_p \{\lambda_1 * L_{match}\left(M'_p, M_p\right)$$

$$+ \lambda_2 * \mathbb{I}_{\{M_p=1\}} \frac{1}{N} \sum_t L_{cls}\left(C'_{p,t}, C_{p,t}\right) \qquad (5)$$

$$+ \lambda_3 * \mathbb{I}_{\{M_p=1,C_t=1\}} \frac{1}{N_{pos}} \sum_t L_{reg}\left(O'_{p,t}, O_{p,t}\right)\}$$

where L_{match} is the cross-entropy loss for the tube-description matching module, L_{cls} and L_{reg} is a cross-entropy loss and IoU loss (i.e., $-ln(\frac{O'_{p,t} \cap O_{p,t}}{O'_{p,t} \cup O_{p,t}})$) for the classification subset and the boundary regression subset, respectively. N and N_{pos} denote the number of frames and the *positive* frames in the tube T_p. It is worth noting that, except for the tube proposals generation network, the rest of the proposed modules can be trained in an end-to-end fashion. However, the loss L_{cls} and L_{reg} in Eq. 5 are only computed for the positive tubes (i.e., $I_{(M_p=1)}$) and positive frames ($I_{(M_p=1,C_{p,t}=1)}$).

F. Inference

During the inference process, given a query sentence s and a video v, our proposed method: **(1)** first generates P tube proposals T_p $(p = 1, 2, ..., P)$. Each tube proposal T_p and the query sentence s form a pair, which will be fed into the visual transformer. **(2)** For each pair of T_p and s, the visual transformer extracts cross-modal features, f_p^{global} and $f_{p,t}^{frame}$. **(3)** The tube-description matching module takes the global feature f_p^{global} as the input and predicts a matching score M'_p. The tube with the highest matching score M'_p is selected and further trimmed. **(4)** For all frames in the selected tube T_p, the classification subnet and the regression subnet take the frame feature $f_{p,t}^{frame}$ as the input, and output the classification score $C'_{p,t}$ and $O'_{p,t} = (\delta l, \delta r)$ for each frame. We follow the method DEBUG [39] to determine the temporal boundary of the selected tube. Specifically, we start from the frame with the highest $C'_{p,t}$. Given its predicted regression offsets $O'_{p,t} = (\delta_l, \delta_r)$, we can produce an initial range $R_{init} = (t - \delta_l * N, t + \delta_r * N)$. Next, we compute R_t for the remaining $N - 1$ frames. If the predicted range of the t-th frame, R_t, has overlap with R_{init}, then R_t is merged with R_{init}. This process is repeated for all frames whose $C'_{p,t}$ higher than a pre-defined threshold ϵ, which will produce the final predicted tube.

V. EXPERIMENTS

A. Evaluation Metrics & Implementation Details

The main evaluation metric for HC-STVG is ***vIoU***, and it is defined as follow:

$$vIoU = \frac{1}{|T_{union}|} \sum_{t \in T_{inter}} IoU\left(B'_t, B_t\right), \qquad (6)$$

where $T_{inter} = T_p \cap T_{GT}$ refers to the intersected frames between the predicted tube T_p and the ground-truth (GT) tube T_{GT}, $T_{union} = T_p \cup T_{GT}$ represents the union set of the predicted frames and the GT frames, B'_t and B_t are the predicted bounding box and the ground truth bounding box at frame t. $vIoU$ reflects the accuracy of the predicted spatio-temporal tubes. The ***vIoU@perc.*** score stands for the

percentage of predicted tubes whose $vIoU$ is larger than *perc.* and $\mathbf{m_{vIoU}}$ is the average of $vIoU$s score over the whole test set.

During the inference process, we sample one frame from every six frames of the tube before feeding the sampled frames into the visual transformer in order to reduce redundancy. The maximum number of bounding boxes detected in each image is set as 101. We use Mask R-CNN pretrained on the Visual Genome dataset [40] to extract the visual features from each bounding box. Every query sentence s is truncated or filled to the maximum length of 40 words. The loss weights λ_1, λ_2, λ_3 are empirically set as 1, 1 and 2, respectively. During the training process, we use Adam optimizer [41]. The initial learning rate and batch size are empirically set as 2e-5 and 32, respectively.

TABLE II
PERFORMANCE (%) OVER m_vIoU, $vIoU@0.3$ AND $vIoU@0.5$ ON THE
VIDSTG DATASET OF THE DECLARATIVE SENTENCE.

Method	Declarative Sentence		
	m_vIoU	vIoU@0.3	vIoU@0.5
Random	0.69%	0.04%	0.01%
GroundeR [42] + TALL [3]	9.78%	11.04%	4.09%
STPR [43] + TALL [3]	10.40%	12.38%	4.27%
WSSTG + TALL [3]	11.36%	14.63%	5.91%
GroundeR [42] + L-Net [44]	11.89%	15.32%	5.45%
STPR [43] + L-Net [44]	12.93%	16.27%	5.68%
WSSTG [13] + L-Net [44]	14.45%	18.00%	7.89%
STGRN [33]	19.75%	25.77%	14.60%
STGVT	**21.62%**	**29.80%**	**18.94%**

TABLE III
PERFORMANCE (%) OF DIFFERENT METHODS ON THE HC-STVG
DATASET IN TERM OF m_vIoU, $vIoU@0.3$ AND $vIoU@0.5$

Method	m_vIoU	vIoU@0.3	vIoU@0.5
Random	0.71%	0.03%	0%
TALL [3]+WSSTG [13]	13.37%	19.95%	7.33%
2D-TAN [45]+WSSTG [13]	15.43%	19.83%	6.81%
STGVT (*w/o trimming*)	16.93%	21.29%	6.64%
STGVT (*w/o pretraining*)	17.46%	25.09%	7.76%
STGVT	**18.15**%	**26.81**%	**9.48**%

B. Baseline Methods

Considering that the existing methods cannot be directly applied for the human-centric spatio-temporal video grounding task, we form two baselines by combining the existing methods on video grounding. Specifically, **TALL** [3] and **2D-TAN** [45] are the two methods for temporal video localization, while **WSSTG** [13] method is proposed for localizing spatio-temporal tubes in *trimmed* videos. So we first apply TALL or 2D-TAN to produce a temporary segment from each *untrimmed* video based on each given sentence, and then utilize the WSSTG method to localize spatio-temporal tubes in the generated segment. The **Random** method localizes the temporal segment and spatial regions randomly. It is worth noting that we replaced the tube-description matching module with the WSSTG method, while keeping the entire training process fully supervised. For a fair comparison, we the same strategy for all these methods to generate the tube proposals, as introduced in Section IV-A.

C. Experimental Results

We report the results of our STGVT method and the baseline methods in Table III. One can observe that our STGVT method outperforms all baseline methods. Our method achieves the m_{vIou} score of 18.15%, which outperforms baseline methods by a significant margin. Our method also surpasses the baselines in terms of $vIoU@0.3$ and $vIoU@0.5$.

To further verify the effectiveness of our method, we also conducted experiments on VidSTG [33] dataset. VidSTG is a suitable dataset for Spatio-Temporal Video Grounding task, but it is not human-centric. Thus we considered all the categories in the library when generating tubes, and use the same method to select the target tube and trim it. Experimental results in Table II. show that our method is also effective on different databases.

D. Ablation Study

1) Effectiveness of Tube-description Matching Module

In our proposed method, the tube-description matching module outputs a raw prediction of the spatio-temporal tube, which will further be trimmed by the tube trimming module. We evaluate the quality of the raw predictions by comparing our method with the WSSTG [13] method in Table IV. It is worth noting that (1) we directly use the tube without any temporal trimming as the final prediction, (2) both ours and WSSTG use the same method for generating tube proposals.

TABLE IV
TUBE MATCHING ACCURACY RESULTS ON THE HC-STVG DATASET

Method	m_vIoU	$vIoU@0.3$	$vIoU@0.5$
Random	2.40%	1.72%	0.44%
WSSTG [13]	12.96%	16.23%	4.35%
STGVT	**16.93%**	**21.29%**	**6.64%**

2) Effectiveness of Tube Trimming Module

In this section, we demonstrate the effectiveness of the tube trimming module in our proposed method. As shown in Table III, our full method (*STGVT*) significantly outperforms the simplified version without the tube trimming module (termed as *STGVT (w/o trimming)*). The performance improvement of our *STGVT* over *STGVT (w/o trimming)* indicates the effectiveness of the tube trimming module. However, it is worth noting that, even without using the tube trimming module, our simplified version *STGVT (w/o trimming)* method still achieves better performance compared with other baseline methods. Besides, we compared the results of the trimmed positive tube, as shown in Table V, the result of tube selection we chose significantly outperforms the baseline method.

TABLE V
TEMPORAL GROUNDING PERFORMANCE (%) OVER m_tIoU AND
m_vIoU

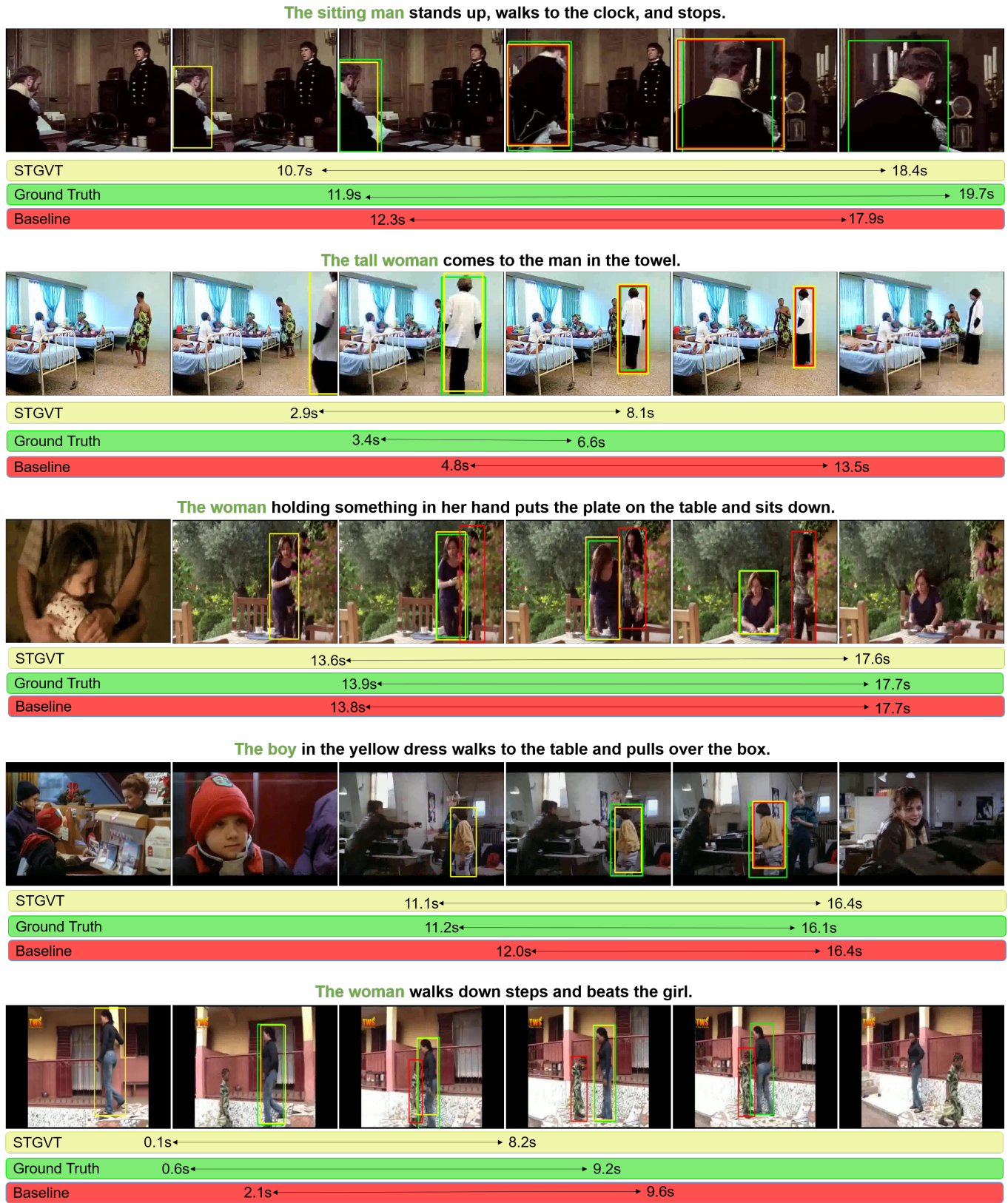| Method	m_vIoU	m_tIoU
Random	0.71%	10.50%
TALL [3] $+Tube_{pos}$	31.64%	44.31%
2D-TAN [45]+$Tube_{pos}$	32.89%	46.20%
STGVT+$Tube_{pos}$	**34.71**%	**48.64**%

Fig. 5. Representative results produced by our method and the baseline method (WSSTG+TALL). In the 2nd and the 5th examples, our method achieves better temporal localization results when compared with the baseline method. While in the 1st, the 3rd, and the 5th examples, our method achieves more accurate spatial localization results.

3) Effectiveness of the Pretraining Strategy

Since we adopt the visual transformer *pretrained* based on the Conceptual Captions dataset [46], we wonder whether the performance improvement is due to the pretraining strategy. So we perform another experiment by training the visual transformer from scratch[2] (and our method is referred as *STGVT (w/o pretraining)*. As shown in Table III, we observe that: 1) our method *STGVT (w/o pretraining)* outperforms existing baseline methods; 2) using the pretraining strategy can improve the performance.

E. Qualitative Analysis

We provide some visualization results in Fig. 5. These examples show that our method outperforms the baseline method for video grounding both spatially and temporally. Specifically, in most examples (except the 3rd one), there are more overlap between our temporal tube prediction results (see the yellow line) and the ground-truth (see the green line) when compared with the baseline method TALL [3] + WSSTG [13](see the red line), which indicates our method can localize the video segment more precisely. In the 3rd example, both our method and the baseline method achieve good performance in temporal grounding, but our method generates more accurate bounding boxes.

VI. CONCLUSION

In this work, we have introduced a novel task Human-centric Spatio-Temporal Video Grounding (HC-STVG). Furthermore, we have contributed a new HC-STVG dataset with rich spatio-temporal tube annotations for videos and the corresponding descriptive sentences. We have also proposed a baseline method named STGVT, which takes advantage of the Visual Transformers to tackle the HC-STVG task. Comprehensive experiments have demonstrated that our method outperforms the existing grounding methods for the HC-STVG task.

REFERENCES

[1] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–85. 1, 2

[2] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798. 1, 2

[3] 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. 2, 3, 4, 8, 10

[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. 2, 3, 4

[5] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8971–8980. 3, 4

[6] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20. 2

[7] S. Huang, S. Liu, T. Hui, J. Han, B. Li, J. Feng, and S. Yan, "Ordnet: Capturing omni-range dependencies for scene parsing," *IEEE Transactions on Image Processing*, vol. 29, pp. 8251–8263, 2020. 2

[8] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, "Referring image segmentation via cross-modal progressive comprehension," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 488–10 497. 2

[9] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, and J. Han, "Linguistic structure guided context modeling for referring image segmentation," *arXiv preprint arXiv:2010.00515*, 2020. 2

[10] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[11] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315. 2

[12] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, and B. Li, "A real-time cross-modality correlation filtering method for referring expression comprehension," *arXiv preprint arXiv:1909.07072*, 2019. 2

[13] Z. Chen, L. Ma, W. Luo, and K.-Y. K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," *arXiv preprint arXiv:1906.02549*, 2019. 2, 4, 5, 8, 10

[14] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," *arXiv: Computation and Language*, 2019. 2, 3

[15] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv: Computer Vision and Pattern Recognition*, 2019. 2, 3

[16] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." in *AAAI*, 2020, pp. 11 336–11 344. 2, 3

[17] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019. 2, 3

[18] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019. 2, 3

[19] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13–23. 2, 3, 6

[20] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019. 2, 3

[21] 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. 3

[22] 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. 3

[23] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9062–9069. 3

[24] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 162–171. 3

[25] S. Chen and Y.-G. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8199–8206. 3

[2]The visual transformer has two branches, one for textual feature while the other for visual feature. All methods except for the baseline method 'Random' use a pretrained transformer for the textual branch.

[26] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 245–253. 3

[27] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257. 3

[28] Y. Yuan, L. Ma, and W. Zhu, "Sentence specified dynamic video thumbnail generation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2332–2340. 3

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 3

[30] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018. 4

[31] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele, "Coherent multi-sentence video description with variable level of detail," in *German conference on pattern recognition*. Springer, 2014, pp. 184–195. 4

[32] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715. 4

[33] 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 conference on computer vision and pattern recognition*, 2020. 4, 5, 8

[34] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, 2013. 4

[35] X. Shang, D. Di, J. Xiao, Y. Cao, and T. S. Chua, "Annotating objects and relations in user-generated videos," in *the 2019*, 2019. 5

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99. 5

[37] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4405–4413. 5, 6

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 6

[39] C. Lu, L. Chen, C. Tan, X. Li, and J. Xiao, "Debug: A dense bottom-up grounding approach for natural language video localization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5147–5156. 7

[40] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017. 8

[41] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011. 8

[42] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290. 8

[43] 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. 8

[44] 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, 2019, pp. 8175–8182. 8

[45] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," *arXiv preprint arXiv:1912.03590*, 2019. 8

[46] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565. 10

Zongheng Tang is currently a PhD candidate in School of Computer Science and Engineering, Beihang University. His current research interests include referring expression and video grounding. He is the co-organizer of CVPR 2021 Person in Context Workshop/challenge.

Yue Liao is currently a PhD candidate in School of Computer Science and Engineering, Beihang University. His current research interests include computer vision and deep learning. He has authorized and co-authorized on more than 10 papers in top-tier academic journals and conferences. He is the co-organizer of ECCV 2018 and ICCV 2019 Person in Context Workshop/challenge. He has been serving as a reviewer for numerous academic conferences such as CVPR, ICCV, ACM MM and ACCV.

Si Liu is an associate professor in Beihang University. She received PHD degree from Institute of Automation, Chinese Academy of Sciences. She has been Research Assistant and Postdoc in National University of Singapore. She was a visiting scholar of Microsoft Research Asia. She has published over 40 cutting-edge papers on image editing and segmentation on TPAMI, IJCV, CVPR, ECCV and ICCV. She was the recipient of Best Paper of ACM MM 2013, Best demo award of ACM MM 2012. She was the Champion of CVPR 2017 Look Into Person Challenge. She is the organizer of ECCV 2018 and ICCV 2019 Person in Context Workshop/challenge. She servers as an area chair of ICCV 2019, CVPR 2020 and ECCV 2020, SPC of IJCAI 2019 and AAAI 2019.

Guanbin Li is currently an associate professor in School of Data and Computer Science, Sun Yat-sen University. He received his PhD degree from the University of Hong Kong in 2016. His current research interests include computer vision, image processing, and deep learning. He is a recipient of ICCV 2019 Best Paper Nomination Award. He has authorized and co-authorized on more than 50 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as TPAMI, TIP, TMM, TC, CVPR, AAAI and IJCAI.

Xiaojie Jin is working as a Research Scientist in Bytedance Inc. in the USA. He received his Ph.D degree from National University of Singapore in 2018. His current research interests include intelligent video editing, multi-modal information processing, neural architecture search, deep model compression and acceleration. His research works have been published in top-tier conferences/journals including ICCV/ECCV/CVPR/ICML/NIPS/ICLR, *etc.* He won two championship prizes in the prestigious ILSVRC competitions in 2015 and 2017, respectively.
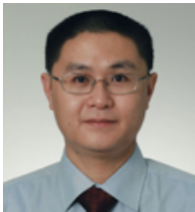
Hongxu Jiang received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 1998 and 2003, respectively. He is currently an Associate Professor with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University. His research interests include image processing, real-time signal processing systems, and embedded systems.
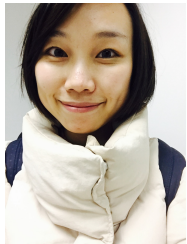
Qian Yu is an associate professor at Beihang University. Before joining Beihang, she was a postdoc at UC Berkeley and International Computer Science Institute, working with Professor Stella Yu. She received her doctorate from the Queen Mary University of London, co-supervised by Professor Yi-Zhe Song and Professor Tao Xiang. Her research interest is sketch understanding and cross-modality modeling. Her research works have been published in top-tier journals and conferences including IJCV, CVPR, ECCV and ICCV. She was the recipient of Best Scientific Paper of BMVC 2015. She is the organizer of ICCV 2021 Sketching for Human Expressivity Workshop.

Dong Xu received the BE and PhD degrees from University of Science and Technology of China, in 2001 and 2005, respectively. While pursuing the PhD degree, he was an intern with Microsoft Research Asia, Beijing, China, and a research assistant with the Chinese University of Hong Kong, Shatin, Hong Kong, for more than two years. He was a postdoctoral research scientist with Columbia University, New York, NY, for one year. He worked as a faculty member with Nanyang Technological University, Singapore. Currently, he is a professor and chair in Computer Engineering with the School of Electrical and Information Engineering, the University of Sydney, Australia. His current research interests include computer vision, statistical learning, and multimedia content analysis. He was the co-author of a paper that won the Best Student Paper award in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2010, and a paper that won the Prize Paper award in IEEE Transactions on Multimedia (T-MM) in 2014. He is a fellow of the IEEE.