Context-Guided Spatio-Temporal Video Grounding

Xin Gu[1,3*] Heng Fan[2*] Yan Huang[2] Tiejian Luo[1] Libo Zhang[1,3†]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Department of Computer Science and Engineering, University of North Texas, Denton, USA
[3]Institute of Software, Chinese Academy of Sciences, Beijing, China

Abstract

Spatio-temporal video grounding (or STVG) task aims at locating a spatio-temporal tube for a specific instance given a text query. Despite advancements, current methods easily suffer the distractors or heavy object appearance variations in videos due to insufficient object information from the text, leading to degradation. Addressing this, we propose a novel framework, context-guided STVG (CG-STVG), which mines discriminative instance context for object in videos and applies it as a supplementary guidance for target localization. The key of CG-STVG lies in two specially designed modules, including instance context generation (ICG), which focuses on discovering visual context information (in both appearance and motion) of the instance, and instance context refinement (ICR), which aims to improve the instance context from ICG by eliminating irrelevant or even harmful information from the context. During grounding, ICG, together with ICR, are deployed at each decoding stage of a Transformer architecture for instance context learning. Particularly, instance context learned from one decoding stage is fed to the next stage, and leveraged as a guidance containing rich and discriminative object feature to enhance the target-awareness in decoding feature, which conversely benefits generating better new instance context to improve localization finally. Compared to existing methods, CG-STVG enjoys object information in text query and guidance from mined instance visual context for more accurate target localization. In experiments on HCSTVG-v1/-v2 and VidSTG, CG-STVG sets new state-of-the-arts in m_tIoU and m_vIoU on all of them, showing efficacy. Code is released at https://github.com/HengLan/CGSTVG.

1. Introduction

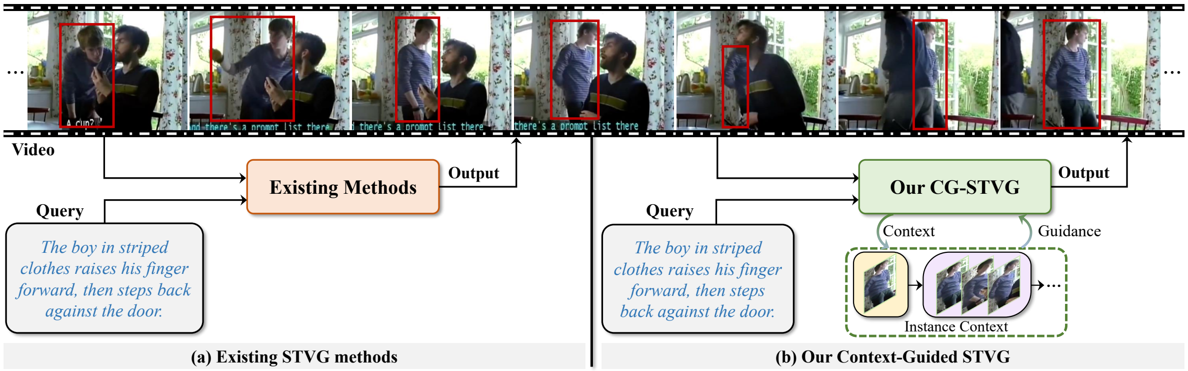
Spatio-temporal video grounding task, or ***STVG***, is recently introduced in [41] and aims to localize the object of interest in an untrimmed video with a spatio-temporal tube (formed by a sequence of bounding boxes) given a *free-form* textual query. It is a challenging multimodal task which is involved with learning and understanding spatio-temporal visual representations in videos and their connections to the linguistic representation of text. Due to the importance in multimodal video understanding, STVG has drawn increasing attention in recent years (*e.g.*, [16, 21, 29, 31, 35, 40, 41]).

Current methods usually use the given textual expression as the *only* cue for retrieving object in videos (see Fig. 1 (a)). Despite progress, they may degrade in complex scenes (*e.g.*, in presence of distractors, or severe appearance changes, or both in videos), because text query is *insufficient* to describe and distinguish the foreground object in these cases. To alleviate this problem, one straightforward solution is to enhance the textual query by including more fine-grained linguistic description. However, there may exist several issues. *First*, this needs reconstruction of text queries for all objects with longer detailed descriptions, which is laborious as well as expensive. *Second*, longer text query will result in more computational overheads for training and inference. *Third*, although the text query can be enhanced with more details, it might still be hard to comprehensively describe certain visual details [43]. Thus, it is natural to ask: *Is there any other way, besides enhancing text query, that improves efficiently, effectively, and friendly spatio-temporal video ground?*

We answer *yes*! Instead of enhancing the text query, we propose to exploit *visual information* of the object to offer a guidance, directly from the *vision perspective*, for improving STVG. As indicated in the famous saying, "*A Picture Is Worth a Thousand Words*", visual cues can provide richer information with description granularity about the target object. Nevertheless, for the STVG task, there is *no* additional *external* visual information allowed, besides the text query, for target localization. So, *where to acquire the desired visual information for improving STVG?*

From the video itself! In this paper, we introduce a novel framework, context-guided STVG or CG-STVG, that mines *internally* discriminative visual context information from a

[*] Equal contributions [†] Corresponding author: libo@iscas.ac.cn


Video
Existing Methods
Output
Query
The boy in striped clothes raises his finger forward, then steps back against the door.
Query
The boy in striped clothes raises his finger forward, then steps back against the door.
Our CG-STVG
Output
Context
Guidance
Instance Context

(a) Existing STVG methods

(b) Our Context-Guided STVG

Figure 1. Comparison between (a) existing methods that localize the target using object information from text query and (b) our CG-STVG that enjoys object information from text query and guidance from mined instance context for STVG. *Best viewed in color for all figures.*

video for the object, and uses it as a supplementary guidance to improve target localization (see Fig. 1 (b)). The crux of CG-STVG lies in two crucial modules, including instance context generation (or ICG) and instance context refinement (or ICR). ICG focuses on discovering visual information of the object. Specifically, ICG first estimates potential regions for the foreground and then uses them to extract contextual information of both appearance and motion from the visual features. Considering there might exist noises in contextual features that are irrelevant or even harmful for the localization due to inaccurate foreground region estimation, ICR is leveraged to eliminate the useless information. Concretely, it adopts a joint temporal-spatio filtering way based on the temporal and spatio relevance scores to suppress irrelevant features, greatly enhancing the context for localization. In this work, we adopt DETR-similar architecture [4] to implement CG-STVG. During video grounding, ICG, together with the ICR, are deployed at each of the decoding stage for instance context learning. Particularly, the instance context learned from one decoding stage is fed to the next stage, and used as a supplementary guidance containing rich and discriminative object information to enhance target-awareness of decoding feature, which in turn benefits generating better new instance context for improving the localization finally. Fig. 2 illustrates the architecture of CG-STVG.

To our best knowledge, CG-STVG is the first to mine instance visual context from the videos to guide STVG. Compared with existing approaches, CG-STVG can leverage the object information from both text query, as in current methods, and guidance from its mined instance context for more accurate target localization. To validate its effectiveness, we conduct extensive experiments on three datasets, including HCSTVG-v1/-v2 [31] and VidSTG [42], CG-STVG outperforms existing methods and sets new state-of-the-arts in m_tIoU and m_vIoU on all of these benchmarks, evidencing the efficacy of guidance from instance context for STVG.

In summary, the main contributions are as follows:

♠ *We introduce CG-STVG, a novel and simple approach for improving STVG via mining instance visual context from the video to guide target localization.*

♡ *We propose an instance context generation module (ICG) to discover visual context information of the object.*

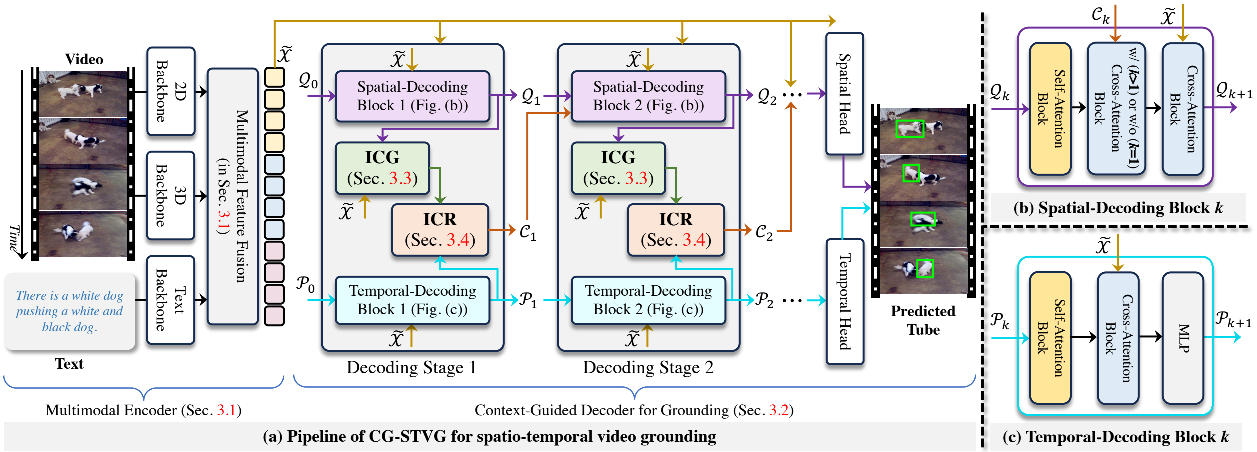
♣ *An instance context refinement (ICR) module is presented to improve the context of object by eliminating irrelevant contextual features, greatly enhancing the performance.*

♢ *In extensive experiments on three benchmarks, including HCSTVG-v1/-v2 [31] and VidSTG [42], CG-STVG sets new state-of-the-arts, showing the effectiveness.*

2. Related Work

Spatio-temporal video grounding. Spatio-temporal video grounding [31] aims to generate a spatio-temporal tube for a target given its text query. Early methods (*e.g.*, [31, 40, 41]) mainly follow a two-stage paradigm, which leverages a pre-trained detector to obtain the candidate region proposals and then finds the correct region proposals through the designed network. The main issue of these methods is the heavy reliance on pre-trained detectors, and the performance is restricted by a detector's own limitations. Differently, recent works (*e.g.*, [16, 21, 29, 35]) adopt a one-stage paradigm, directly generating spatio-temporal object proposals without relying on any pre-trained object detectors. The method of [29] is the first of this kind, which leverages the visual-linguistic transformer to generate a spatio-temporal object tube corresponding to the textual sentence. Inspired by the success of the model for text-conditioned object detection [17], the method in [35] introduces a spatio-temporal transformer decoder together with a video-text encoder for STVG. The approach of [16] utilizes a multi-modal template as the global objective to deal with the inconsistency issue for improvement. The work of [21] proposes to explore static appearance and dynamic motion cues collaboratively for target localization, showing promising results.

In this paper, we exploit visual context from videos and



(a) Pipeline of CG-STVG for spatio-temporal video grounding

(b) Spatial-Decoding Block k

(c) Temporal-Decoding Block k

Figure 2. Overview of our method, which consists of a multimodal encoder for feature extraction and a context-guided decoder by cascading a set of decoding stages for grounding. In each decoding stage, instance context is mined to guide query learning for better localization.

adopt it as a guidance for target localization. ***Different*** from existing approaches (*e.g.*, [16, 21, 29, 35]) which explore object information only from the text query for localization, the proposed CG-STVG is able to leverage both textual cue and object guidance from the mined instance context, significantly enhancing the STVG performance and outperforming other methods, particularly in complicated scenarios with similar distractors or large appearance changes.

Temporal Grounding. Temporal grounding aims at locating and understanding specific objects or events in a video. Relevant to but different than the STVG, temporal grounding does not require bounding box localization of the target. Numerous approaches (*e.g.*, [2, 3, 6, 10, 24, 33, 39]) have been introduced recently. For example, the algorithm of [2] proposes an effective strategy to avoid the long-form burden by applying a guidance model for grounding time. The approach of [3] leverages cross-modal contrastive learning at coarse-grained (video-sentence) and fine-grained (clip-word) levels for grounding. The work in [6] designs a multimodal framework to learn complementary features from images, flow, and depth for the temporal grounding. ***Different*** than these methods, we focus on the more challenging STVG that spatially and temporally localizes the object.

Vision-Language Modeling. Vision-language modeling is to simultaneously process visual and linguistic information for joint multimodal understanding in various tasks such as visual question answering [1, 5, 15, 19, 27, 38], video captioning [7, 13, 26, 28, 36, 44], text-to-image generation [20, 25], visual-language tracking [8, 45], refer video segmentation [9], etc. ***Differently***, we focus on modeling vision and language for spatio-temporal target localization.

3. The Proposed Method

Overview. In this work, we present CG-STVG to mine discriminative visual context of object and use it as a guidance to improve localization. Inspired by DETR [4], CG-STVG employs an encoder-decoder architecture, which comprises a multimodal encoder (Sec. 3.1) and a context-guided decoder (Sec. 3.2). As in Fig. 2, the encoder aims at generating multimodal visual-linguistic feature that contains object information from text query, which is sent to the context-guided decoder for target localization guided by instance context learned with ICG (Sec. 3.3) and ICR (Sec. 3.4).

3.1. Multimodal Encoder

The multimodal encoder is to generate a robust multimodal feature for the target localization in decoder, and consists of visual and textual feature extraction and fusion as follows.

Visual Feature Extraction. To leverage rich cues from the videos, we extract both the appearance and motion features. In specific, we first sample a set of frames $\mathcal{F} = \{f_i\}_{i=1}^{N_v}$ of length N_v from the video, and then utilize ResNet-101 [11] for appearance feature extraction and VidSwin [23] for motion feature extraction, respectively. We denote the appearance feature as $\mathcal{V}_a = \{v_i^a\}_{i=1}^{N_v}$, where $v_i^a \in \mathbb{R}^{H\times W\times C_a}$ with H, W, and C_a the height, width and channel dimensions. Similarly, we denote the motion feature as $\mathcal{V}_m = \{v_i^m\}_{i=1}^{N_v}$, where $v_i^m \in \mathbb{R}^{H\times W\times C_m}$ with C_m the channel dimension.

Textual Feature Extraction. We adopt RoBERTa [22] for textual feature extraction. We first tokenize query to obtain a word sequence $\mathcal{W} = \{w_i\}_{i=1}^{i=N_t}$ and then apply RoBERTa to produce an embedding sequence $\mathcal{T} = \{t_i\}_{i=1}^{i=N_t}$, where $t_i \in \mathbb{R}^{C_t}$ with C_t the word embedding dimension.

Multimodal Feature Fusion. STVG is a multimodal task. To enhance feature representation, we perform multimodal fusion of the appearance feature $\mathcal{V}_a$, motion feature $\mathcal{V}_m$, and text feature $\mathcal{T}$. Specifically, we first map $\mathcal{V}_a$, $\mathcal{V}_m$ and $\mathcal{T}$ to the same channel number through linear projection and then concatenate corresponding features to obtain the representation of multimodal features $\mathcal{X} = \{x_i\}_{i=1}^{N_v}$ as follows,

$$x_i = [\underbrace{v_{i_1}^a, v_{i_2}^a, ..., v_{i_{H\times W}}^a}_{\text{appearance features } v_i^a}, \underbrace{v_{i_1}^m, v_{i_2}^m, ..., v_{i_{H\times W}}^m}_{\text{motion features } v_i^m}, \underbrace{t_1, t_2, ..., t_N}_{\text{textual features}}$$

where x_i is the multimodal feature in frame i. Then, we ac position embedding $\mathcal{E}_{pos}$ and type embedding $\mathcal{E}_{typ}$ to $\mathcal{X}$ t

$$\mathcal{X}' = \mathcal{X} + \mathcal{E}_{pos} + \mathcal{E}_{typ}$$

Finally, we perform multimodal feature fusion by applyir a self-attention encoder on $\mathcal{X}'$ as follows,

$$\tilde{\mathcal{X}} = \texttt{SAEncoder}(\mathcal{X}')$$

where $\tilde{\mathcal{X}}$ is the enhanced multimodal feature for decodin and $\texttt{SAEncoder}(\cdot)$ the self-attention encoder with L (L= standard self-attention encoder blocks [32]. Please refer **supplementary material** for architecture of $\texttt{SAEncoder}(\cdot)$.

3.2. Context-Guided Decoder for Grounding

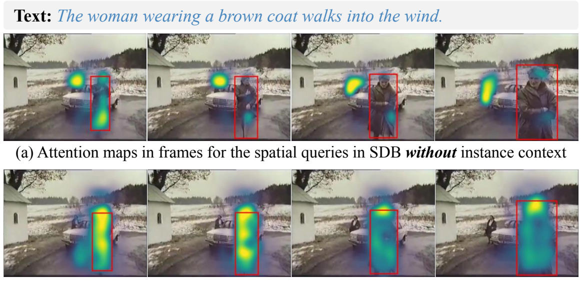
CG-STVG designs a context-guided decoder with K stages in a cascade for grounding as in Fig. 2 (a). Since CG-STVG needs to locate target spatially and temporally, each decoding stage has two blocks, including a *spatial-decoding block* (SDB) and a *temporal-decoding block* (TDB), for spatial and temporal feature learning. In each stage (except for the first), instance context by ICG and ICR (see later) is applied as a guidance with rich visual cue to enhance the query feature, which is in turn used to generate new instance context.

Specifically, let $\mathcal{Q}_{k-1}$=$\{q_i^{k-1}\}_{i=1}^{N_v}$ denote spatial query features for N_v frames and $\mathcal{P}_{k-1}$=$\{p_i^{k-1}\}_{i=1}^{N_v}$the temporal query features sent to the k^{th} ($1 < k \leq K$) decoding stage. $\mathcal{Q}_0$ and $\mathcal{P}_0$ fed to the first decoding stage are initialized following DETR [4]. Then, in decoding stage k, we use SDB$_k$ to learn query feature $\mathcal{Q}_k$ using instance context $\mathcal{C}_{k-1}$ from decoding stage (k-1) as a guidance and multimodal feature $\tilde{\mathcal{X}}$ from the encoder. As in Fig. 2 (b), SDB$_k$ contains three components with one self-attention and two cross-attention blocks. The self-attention block is to enhance query features by interacting them. The former cross-attention block aims to guide query features using $\mathcal{C}_{k-1}$, while the later is for learning object position information from $\tilde{\mathcal{X}}$. The process of SDB$_k$ for learning $\mathcal{Q}_k$ can be formulated as follows,

$$\begin{aligned}\mathcal{Q}_k &= \texttt{SDB}_k(\mathcal{Q}_{k-1}, \mathcal{C}_{k-1}, \tilde{\mathcal{X}}) \\ &= \texttt{CA}(\texttt{CA}(\texttt{SA}(\mathcal{Q}_{k-1}), \mathcal{C}_{k-1}), \tilde{\mathcal{X}})\end{aligned}$$

where $\texttt{SA}(\mathbf{z})$ denotes the self-attention block with $\mathbf{z}$ generating query/key/value, and $\texttt{CA}(\mathbf{z}, \mathbf{u})$ the cross-attention block with $\mathbf{z}$ generating query and $\mathbf{u}$ key/value, as in [32]. Due to limited space, please see **supplementary material** for detailed architectures. For SDB$_1$, because the instance context does not exist, $\mathcal{Q}_1$ is learned as follows,

$$\mathcal{Q}_1 = \texttt{SDB}_1(\mathcal{Q}_0, \tilde{\mathcal{X}}) = \texttt{CA}(\texttt{SA}(\mathcal{Q}_0), \tilde{\mathcal{X}})$$


Text: *The woman wearing a brown coat walks into the wind.*

(a) Attention maps in frames for the spatial queries in SDB ***without*** instance context

(b) Attention maps in frames for the spatial queries in SDB ***with*** instance context

Figure 3. Attention maps for spatial queries in video frames in the spatial-decoding block *without* (image (a)) and *with* our proposed instance context (image (b)). We can clearly see that our instance context effectively improves target-awareness in the spatial queries and thus the target position information learning for localization. The red boxes indicate the foreground object to localize.

In decoding, the spatial query feature aims to learn object information progressively from $\tilde{\mathcal{X}}$. In our SDB, the spatial query feature is guided by the visual context of the object to enhance its *target-awareness* in *vision perspective* such that it can explicitly exploit rich and discriminative visual cues to learn more accurate position information from $\tilde{\mathcal{X}}$ for better target localization, even when text cannot well describe the object, which significantly differs than existing methods (*e.g.*, [16, 21, 29, 35]).

Similarly in decoding stage k, the temporal query feature $\mathcal{P}_k$ is learned by TDB$_k$ which consists of self-attention and cross-attention blocks followed by the MLP, as in Fig. 2 (c). The process for learning $\mathcal{P}_k$ can be expressed as follows,

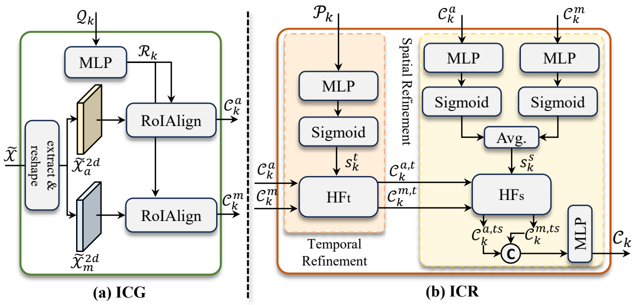
$$\mathcal{P}_k = \texttt{TDB}_k(\mathcal{P}_{k-1}, \tilde{\mathcal{X}}) = \texttt{MLP}(\texttt{CA}(\texttt{SA}(\mathcal{P}_{k-1}), \tilde{\mathcal{X}}))$$

Notice that, instance context $\mathcal{C}_{k-1}$ is not used in TDB, as it mainly works to localize target when it exists in the frames, instead of detecting if the object exists or not. When applying instance context in TDB, it even cause slight performance drop. Thus, instance context is only applied in STB.

Once generating $\mathcal{Q}_k$ and $\mathcal{P}_k$, they are used to learn new instance context $\mathcal{C}_k$ in decoding stage k with already acquired object position and frame information using ICG and ICR (as explained later), which will be applied to guide further query learning in subsequent stages for improving target-awareness and position information learning, as evidenced in Fig. 3, in a progressive way. In the decoding stage K, the learned $\mathcal{Q}_K$ and $\mathcal{P}_K$ are fed to two heads to predict the final object boxes $\mathcal{B}_K = \{b_i\}_{i=1}^{N_v}$, where $b_i \in R^4$ denotes the central position, width and height of the prediction box, and the start and end probabilities of each frame $\mathcal{H}_K = \{(h_i^s, h_i^e)\}_{i=1}^{N_v}$, where the start and end times are determined by the maximum joint start and end probability.

3.3. Instance Context Generation (ICG)

To exploit instance context in the video, we introduce a simple yet effective module, termed instance context generation



(a) ICG

(b) ICR

Figure 4. Illustration ICG (image (a)) and ICR (image (b)).

(ICG). Specifically, ICG is deployed in each decoding stage k of the context-guided decoder and takes the spatial query feature $\mathcal{Q}_k$ learned from SDB$_k$ to discover the potential features of the foreground (*i.e.*, the so-called *instance context*), as in Fig. 4 (a). The **intuition** is that, during the progressive video grounding for localization, $\mathcal{Q}_k$ has gradually learned more and more position information for the object and thus can be employed to find target regions in a video, which are used to further mine object features in the decoding stage k.

To this end, we first utilize a three-layer MLP in ICG to transform $\mathcal{Q}_k$ into foreground regions $\mathcal{R}_k$ as follows,

$$\mathcal{R}_k = \texttt{MLP}(\mathcal{Q}_k) = \{r_i^k\}_{i=1}^{N_v}$$

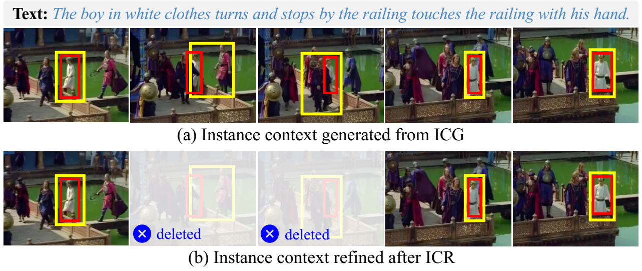
where $r_i^k \in R^4$ denotes estimated object center and scale in frame i. Then with $\mathcal{R}_k$, we leverage it to extract the corresponding foreground features, including both appearance and motion features. In specific, we first extract the appearance and motion features, denoted by $\tilde{\mathcal{X}}_a$ and $\tilde{\mathcal{X}}_m$, from the multimodal feature $\tilde{\mathcal{X}}$, and then reshape them into 2D feature maps $\tilde{\mathcal{X}}_a^{2d}$=reshape($\tilde{\mathcal{X}}_a$) and $\tilde{\mathcal{X}}_m^{2d}$=reshape($\tilde{\mathcal{X}}_m$). After that, we use RoIAlign [12] to extract appearance and motion instance context as follows,

$$\mathcal{C}_k^a = \texttt{RoIAlign}(\tilde{\mathcal{X}}_a^{2d}, \mathcal{R}_k) \quad \mathcal{C}_k^m = \texttt{RoIAlign}(\tilde{\mathcal{X}}_m^{2d}, \mathcal{R}_k)$$

where $\mathcal{C}_k^a$ denotes the appearance instance context and $\mathcal{C}_k^m$ the motion instance context. $\mathcal{C}_k^a$ mainly encompass various rich visual attributes of the target, such as shape, texture and color, while $\mathcal{C}_k^a$ predominantly captures motion properties of the object, including speed and trajectory. Both of these two context are beneficial to enhance the target-awareness, enhancing target-awareness in spatial query feature for better target position learning.

3.4. Instance Context Refinement (ICR)

Considering that the estimated foreground regions may contain noise because the target position information in $\mathcal{Q}_k$ is not enough, the instance visual context of $\mathcal{C}_k^a$ and $\mathcal{C}_k^m$ might contain irrelevant and even harmful features, and thus is degraded. To remedy, we further present the instance context refinement module (ICR) to refine $\mathcal{C}_k^a$ and $\mathcal{C}_k^m$ for better final instance context by eliminating irrelevant features. In particular, we introduce a two-level temporal-spatial joint refinement mechanism in ICR. In the first level, instance context is refined by a temporal filter with temporal-confidence of each feature. Then, at the second level, a spatial filter is designed to suppress irrelevant features. Fig. 4 (b) shows the architecture of ICR, which is detailed as follows.


Text: The boy in white clothes turns and stops by the railing touches the railing with his hand.
deleted
deleted

(a) Instance context generated from ICG

(b) Instance context refined after ICR

Figure 5. Illustration of ICR for context refinement. The red boxes indicate the foreground, while yellow boxes the instance context. We can see that, our ICR is able to help eliminate irrelevant features in the initial instance context generated from ICG.

Temporal Refinement. Because instance context is only related to the object in a certain temporal window, instead of the whole video, we leverage the temporal query feature $\mathcal{P}_k$ in decoding stage k to calculate the confidence score of each frame being relevant to the object. Specifically, we simply apply an MLP module followed by a Sigmoid function to transform $\mathcal{P}_k$ to the temporal confidence scores as follows,

$$s_k^t = \texttt{Sigmoid}(\texttt{MLP}(\mathcal{P}_k))$$

where $s_k^t \in R^{N_v}$ represents the temporal confidence scores. The higher the $s_k^t(i)$ is, the more relevant the instance feature in frame i is. To eliminate irrelevant feature, we design a filter to drop instance context features with temporal confidence scores lower than a preset threshold θ^t as follows,

$$\mathcal{C}_k^{a,t} = \texttt{HF}_\texttt{t}(\mathcal{C}_k^a, s_k^t, \theta^t) \qquad \mathcal{C}_k^{m,t} = \texttt{HF}_\texttt{t}(\mathcal{C}_k^m, s_k^t, \theta^t)$$

where $\mathcal{C}_k^{a,t}$ and $\mathcal{C}_k^{m,t}$ are refined instance context. $\texttt{HF}_t$ is a threshold operation that passes instance context features of $\mathcal{C}_k^a$ and $\mathcal{C}_k^m$ with confidence scores greater than θ^t.

Spatial Refinement. Different from the temporal refinement, spatial refinement aims to measure the quality of context features $\mathcal{C}_k^a$ and $\mathcal{C}_k^m$ from spatial dimension. To this end, we apply two MLP modules with each followed by a Sigmoid function to compute the spatial appearance and motion confidence scores, which are averaged to obtain the final spatial confidence scores, as follows,

$$s_k^s = (\texttt{Sigmoid}(\texttt{MLP}(\mathcal{C}_k^a)) + \texttt{Sigmoid}(\texttt{MLP}(\mathcal{C}_k^m)))/2$$

where $s_k^s \in R^{N_v}$ represents spatial confidence scores which are measured using the predicted IoU confidence [14], originally used for detection. To suppress irrelevant features, we drop features in $\mathcal{C}_k^{a,t}$ and $\mathcal{C}_k^{m,t}$ with spatial confidence scores lower than a preset threshold θ_s as follows,

$$\mathcal{C}_k^{a,ts} = \texttt{HF}_\texttt{s}(\mathcal{C}_k^{a,t}, s_k^s, \theta^s) \qquad \mathcal{C}_k^{m,ts} = \texttt{HF}_\texttt{s}(\mathcal{C}_k^{m,t}, s_k^s, \theta^s)$$

where $\mathcal{C}_k^{a,ts}$ and $\mathcal{C}_k^{m,ts}$ are refined instance context. $\texttt{HF}_s$ is a threshold operation that passes instance context features with confidence scores greater than θ^s. Fig. 5 illustrates the instance refinement by ICR.

Final Instance Context. After the two-level refinement, we concatenate $\mathcal{C}_k^{a,ts}$ and $\mathcal{C}_k^{m,ts}$ and apply an MLP module to obtain the final instance context $\mathcal{C}_k$ as in Fig. 4, which is used to improve target position learning in the next stage.

3.5. Optimization

Given a video and its text, after the k^{th} decoding stage, we predict: (1) start timestamps $\mathcal{H}_k^s = \{h_i^s\}_{i=1}^{N_v}$ and end timestamps $\mathcal{H}_k^e = \{h_i^e\}_{i=1}^{N_v}$ of the video clip related to text, (2) bounding box $\mathcal{B}_k = \{b_i\}_{i=1}^{N_v}$ of the object on which the text focuses, (3) temporal and spatial confidence scores s_k^t and s_k^s in context refinement. In training, we are given groundtruth start timestamps $\mathcal{H}_s^*$, the end timestamps $\mathcal{H}_e^*$, the bounding box sequence $\mathcal{B}^*$.

For temporal grounding, the KL divergence and binary cross-entropy are used as the loss function and the losses of start and end times are computed as follows,

$$\begin{aligned}\mathcal{L}_k^t =& \lambda_s \mathcal{L}_{KL}(\mathcal{H}_s^*, \mathcal{H}_k^s) + \lambda_e \mathcal{L}_{KL}(\mathcal{H}_e^*, \mathcal{H}_k^e) + \\ & \lambda_t \mathcal{L}_{BCE}((\mathcal{H}_s^*, \mathcal{H}_e^*), s_k^t)\end{aligned}$$

For spatial grounding, smooth L1 loss, IoU loss and binary cross-entropy loss are used as follows,

$$\begin{aligned}\mathcal{L}_k^s =& \lambda_l \mathcal{L}_{L_1}(\mathcal{B}^*, \mathcal{B}_k) + \lambda_i \mathcal{L}_{IoU}(\mathcal{B}^*, \mathcal{B}_k) + \\ & \lambda_b \mathcal{L}_{BCE}(IoU(\mathcal{B}^*, \mathcal{B}_k), s_k^s)\end{aligned}$$

The total training loss for training is $\mathcal{L} = \sum_{k=1}^{K}(\mathcal{L}_k^t + \mathcal{L}_k^s)$.

4. Experiments

Implementation. Our CG-STVG is implemented using PyTorch. We use ResNet-101 [11] as 2D backbone, VidSwin-tiny [23] as 3D backbone, and RoBERTa-base [22] as text backbone. Following [16, 21], we utilize pre-trained MDETR [17] to initialize the 2D backbone and text backbone. We use the Adam optimization algorithm [18] with a weight decay of $1e-4$ to end-to-end train our method. The initial learning rate for three backbones is set to $2e-5$ and $3e-4$ for the rest modules. We uniformly resize the video frames to a short side of H=420 and data augmentation methods such as random resizing and random cropping are applied to all training videos. The number of attention heads is set to 8 and the hidden dimension of the encoder and decoder is 256. The batch size is set to 16 in HCSTVG-v1, 32 in HCSTVG-v2 and 64 in VidSTG dataset. The loss weight parameters λ_s, λ_e, λ_t, λ_l, λ_i, λ_b are set to 10, 10, 1, 5, 3, 1, respectively. The number of decoding stages K is set to 6. We set the video frame length N_v to 64 and the text sequence length N_t to 30. The dimensions of the appearance feature, motion feature and text embedding C_a, C_m and C_t are 2048, 768, 768. The temporal threshold θ^t and spatial threshold θ^s are set to 0.7 and 0.8, respectively.

4.1. Datasets and Metrics.

Datasets. Extensive experiments are conducted on three datasets, *i.e.*, HCSTVG-v1 [31], HCSTVG-v2 [31] and VidSTG [41]. HCSTVG, focusing solely on humans in videos, is available in two versions: HCSTVG-v1 and HCSTVG-v2. Following [16, 31, 35], we divide the HCSTVG-v1 into $4,500$ and $1,160$ video-sentence pairs for training and testing, respectively. HCSTVG-v2 further expands HCSTVG-v1, which includes $10,131$, $2,000$, and $4,413$ samples for training, validation, and testing, respectively. As the annotations for test set are not publicly available, we present the results based on validation set as existing methods [21, 35]. VidSTG is another dataset constructed based on video relation dataset. Following [16, 21, 35], VidSTG is divided into training, validation, and test subsets with $80,684$, $8,956$, and $10,303$ distinct sentences, respectively, and $5,436$, 602, and 732 distinct videos, respectively.

Metrics. Following [16, 29, 35], we use m_tIoU, m_vIoU and vIoU@R as evaluation metrics. m_tIoU measures temporal localization performance, while m_vIoU and vIoU@R evaluate spatial localization. In specifc, m_tIoU represents the average tIoU score over all testing sequences and tIoU is calculated as $\frac{|\mathcal{P}_i|}{|\mathcal{P}_u|}$, where $\mathcal{P}_i$ and $\mathcal{P}_u$ represent the intersection and union between the predicted segments and the ground-truth segments, respectively. Similarly, m_vIoU represents the average vIoU score over all testing videos and vIoU is calculated as $\frac{1}{|\mathcal{P}_u|}\sum_{t\in\mathcal{P}_i}\text{IoU}(b_t^*, b_t)$, where b_t^* and b_t are the groundtruth bounding box and the predicted bounding box of the t-th frame. As for vIoU@R, it represents the ratio of samples with vIoU $>$ R in test subset.

4.2. State-of-the-art Comparison

HCSTVG-v1 and HCSTVG-v2. To validate the effectiveness of CG-STVG, we compare it with other state-of-the-arts on HCSTVG-v1 and HCSTVG-v2. Tab. 1 shows the results on the HCSTVG-v1 test set, and our proposed method achieves state-of-the-art performance in 3 out of 4 metrics. Specifically, our method improves the 3.4 absolute m_tIoU score compared to STCAT [16] and improves 1.5 absolute m_vIoU score compared to CSDVL [21]. Compared to our baseline that does not use the proposed instance context by removing ICG and ICR modules, our method achieves improvements of 2.4, 1.9, 2.9, and 4 scores on the four metrics, respectively. On the validation set of the HCSTVG-v2, our method also achieves SOTA in 3 out of 4 metrics as shown

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
STGVT [TCSVT21] [31]	-	18.2	26.8	9.5
STVGBert [ICCV2021] [29]	-	20.4	29.4	11.3
TubeDETR [CVPR22] [35]	43.7	32.4	49.8	23.5
STCAT [NeurIPS22] [16]	49.4	35.1	57.7	30.1
CSDVL [CVPR23] [21]	-	36.9	**62.2**	34.8
Baseline	50.4	36.5	58.6	32.3
CG-STVG	**52.8** (+2.4)	**38.4** (+1.9)	61.5 (+2.9)	**36.3** (+4.0)

Table 1. Comparison with others on HCSTVG-v1 test set (%).

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
PCC [arxiv2021] [37]	-	30.0	-	-
2D-Tan [arxiv2021] [30]	-	30.4	50.4	18.8
MMN [AAAI22] [34]	-	30.3	49.0	25.6
TubeDETR [CVPR22] [35]	-	36.4	58.8	30.6
CSDVL [CVPR23] [21]	58.1	38.7	**65.5**	33.8
Baseline	58.6	37.8	62.4	32.1
CG-STVG	**60.0** (+1.4)	**39.5** (+1.7)	64.5 (+2.1)	**36.3** (+4.2)

Table 2. Comparison with others on HCSTVG-v2 val. set (%).

Methods	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
STGRN [CVPR20] [42]	48.5	19.8	25.8	14.6	47.0	18.3	21.1	12.8
OMRN [IJCAI20] [40]	50.7	23.1	32.6	16.4	49.2	20.6	28.4	14.1
STGVT [TCSVT21] [31]	-	21.6	29.8	18.9	-	-	-	-
STVGBert [ICCV21] [29]	-	24.0	30.9	18.4	-	22.5	26.0	16.0
TubeDETR [CVPR22] [35]	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT [NeurIPS22] [16]	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
CSDVL [CVPR23] [21]	-	33.7	47.2	32.8	-	28.5	39.9	26.2
Baseline	49.7	32.4	45.0	31.4	48.8	27.7	38.7	25.6
CG-STVG	**51.4** (+1.7)	**34.0** (+1.6)	**47.7** (+2.7)	**33.1** (+1.7)	**49.9** (+1.1)	**29.0** (+1.3)	**40.5** (+1.8)	**27.5** (+1.9)

Table 3. Comparison with existing state-of-the-art methods on VidSTG test set (%).

ICG	ICR	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
-	-	50.42	36.52	58.62	32.33
✓	-	51.07	37.42	59.48	32.93
✓	T	51.26	37.86	60.95	33.28
✓	S	52.80	38.04	60.90	35.40
✓	S+T	**52.84**	**38.42**	**61.47**	**36.29**

Table 4. Ablation study of ICG and ICR on HCSTVG-v1 test set of. "T" and "S" represent the temporal and spatial refinement.

θ^t	m_tIoU	m_vIoU	vIoU@0.5
0.3	52.82	38.19	35.34
0.5	52.80	38.29	35.43
0.7	**52.84**	**38.42**	**36.29**
0.9	52.84	38.27	36.12

(a) Ablation study for θ^t.

θ^s	m_tIoU	m_vIoU	vIoU@0.5
0.4	51.64	37.47	32.41
0.6	51.86	37.44	31.64
0.8	**52.84**	**38.42**	**36.29**
0.9	51.79	37.61	32.33

(b) Ablation study for θ^s.

Table 5. Ablation of thresholds in ICR on HCSTVG-v1 test set.

Usage of s_k^t and s_k^s	m_vIoU	vIoU@0.3	vIoU@0.5
Two-level (ours)	**38.42**	**61.47**	**36.29**
one-level w/ "$s_k^t + s_k^s$"	38.31	61.12	35.69
one-level w/ "$s_k^t \times s_k^s$"	38.25	61.07	35.52

Table 6. Ablation of usage of temporal and spatial confidence.

in Tab. 2. CSDVL [21] won the first place in the HCSTVG track of the 4-th Person in Context Challenge. Compared to the CSDVL, our approach outperforms it by 1.9, 0.8 and 2.5 scores on m_tIoU, m_vIoU, and vIoU@0.5 metrics, respectively. The significant improvement in metric vIoU@0.5 across two datasets indicates that instance context excels at refining bounding boxes with an IoU under 0.5.

TDB	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
w/o Instance Context	**52.84**	**38.42**	**61.47**	**36.29**
w/ Instance Context	52.61	38.01	61.03	35.78

Table 7. Ablation on applying instance context to TDB.

VidSTG Dataset. Besides HCSTVG-v1/-v2, we compare CG-STVG with other methods on the challenging VidSTG dataset in Tab. 3. As shown, our method achieves the best results on all 8 metrics for both declarative sentences and interrogative sentences. With the proposed instance context, our method shows an improvement of 1.7 m_tIoU scores and 1.6 m_vIoU scores for declarative sentences and a gain of 1.1 m_tIoU scores and 1.3 m_vIoU scores for interrogative sentences over the baseline. The experimental results further evidence the effectiveness of our method, showing that instance context information helps ground the target.

4.3. Ablation Study

Impact of ICG and ICR. The key of CG-STVG lies in two simple yet effective modules, including ICG and ICR, for instance context learning. To verify their effectiveness, we conduct ablation experiments on HCSTVG-v1 in Tab. 4. As in Tab. 4, our baseline achieves a m_vIoU score of 36.52 without ICG and ICR. After incorporating ICG for instance context, the m_vIoU score is increased to 37.42, demonstrating that the visual context from ICG helps improve the grounding performance. To enhance the quality of instance context, we use a spatial-temporal joint refinement mechanism in ICR module. When we apply temporal refinement alone, we observe that the m_vIoU score is improved by

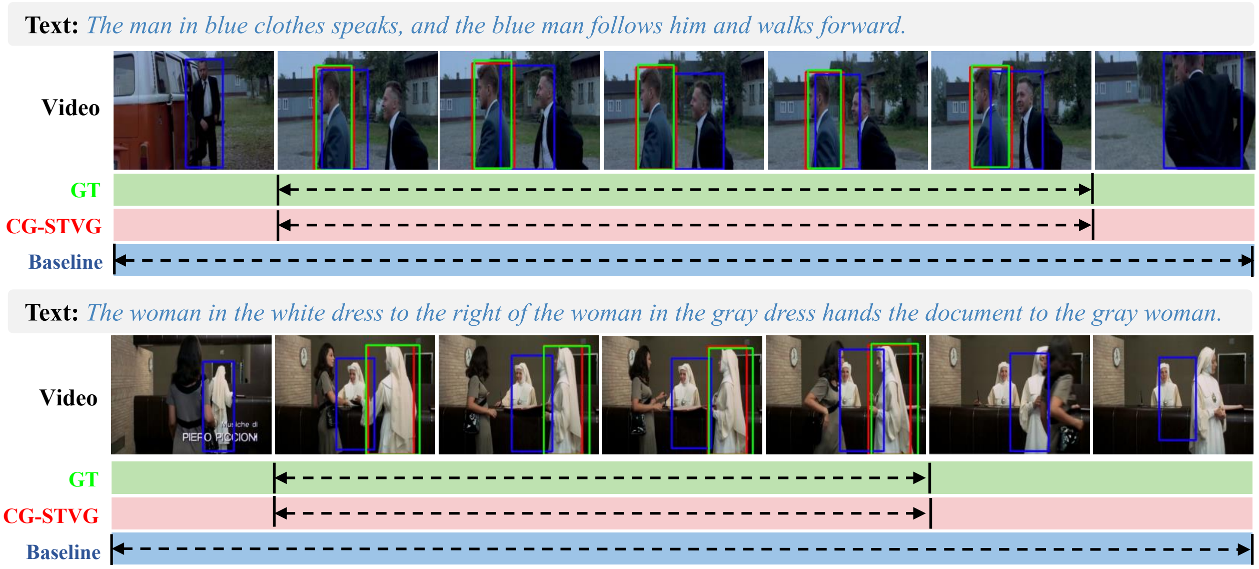

Text: The man in blue clothes speaks, and the blue man follows him and walks forward.
Video
GT
CG-STVG
Baseline
Text: The woman in the white dress to the right of the woman in the gray dress hands the document to the gray woman.
Video
GT
CG-STVG
Baseline

Figure 6. Qualitative results on HCSTVG-v1 test set. Our method (red) shows better localization than the baseline (blue).

0.44. Applying the spatial refinement alone results in a slightly higher increase of 0.62. However, when we use both spatial and temporal refinements simultaneously, the performance improvement is the most significant, with an increase of 1.0, 37.42 vs 38.42. This shows the synergistic effect of spatial and temporal refinements and underscores the effectiveness of our proposed spatial-temporal joint refinement mechanism in enhancing model performance.

Impact of temporal and spatial thresholds in ICR. To improve the quality of the instance context, we use ICR to filter the visual context from ICG. The ICR module refines the instance context through a two-level temporal-spatial joint refinement mechanism. Within this mechanism, there are two crucial parameters, temporal threshold θ^t and spatial threshold θ^s, which are used as standards to filter the context. To investigate the influence of the temporal and spatial threshold on the model, we perform ablation experiments at different thresholds, as shown in Tab. 5. We can see that the model performs best when θ^t is 0.7 and θ^s is 0.8.

Impact of temporal and spatial confidence score usage. Temporal and spatial confidence scores s_k^t and s_k^s are crucial for instance context refinement. In this work, we adopt a two-level method to separately use s_k^t and s_k^s for refinement. To further study the impact of different methods for the usage of temporal and spatial confidence scores, we design two additional one-level methods for refinement: one is to add s_k^t and s_k^s and then apply a single fused confidence for refinement (one-level with "$s_k^t + s_k^s$"), and the other is to multiple s_k^t and s_k^s for refinement (one-level with "$s_k^t \times s_k^s$"). We show the architectures of these two variants and comparison with our strategy in **supplementary material**. We conducted experiments as in Tab. 6, and we can see that our two-level method achieves better performance.

Impact of applying instance context to TDB. From the Tab. 4, it can be seen that as the spatial grounding improves with the help of context, the temporal grounding is also improving, 50.42 vs 52.84. To explore the impact of applying the instance context to the TDB on model performance, we conduct ablation study as shown in Tab. 7. There is a slight drop in model performance after employing context to the TDB. We believe the temporal branch is mainly used to determine the boundaries of events, and the context from the spatial branch has a gap with the temporal branch. Directly using context in temporal branch may cause boundary blur.

4.4. Qualitative Analysis

We present qualitative results in Fig. 6. Compared with our baseline, CG-STVG could accurately locate the target temporally and spatially with instance context.

Due to limited space, we show more results and comparisons as well as analysis in the **supplementary material**.

5. Conclusion

In this work, we present CG-STVG for improving STVG via exploiting instance visual context to guide target localization. The strength of CG-STVG comes from two key modules, including ICG that mines coarse visual context, and ICR that refines this context using time and space information. The experimental results on three benchmarks further demonstrate the effectiveness of our method.

Acknowledgement. Libo Zhang is supported by the Youth Innovation Promotion Association, CAS (2020111). Heng Fan and Yan Huang and their employer receive no financial support for this work.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 3

[2] Wayner Barrios, Mattia Soldan, Fabian Caba Heilbron, Alberto Mario Ceballos-Arroyo, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In *ICCV*, 2023. 3

[3] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *ECCV*, 2022. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3, 4

[5] Long Chen, Yuhang Zheng, and Jun Xiao. Rethinking data augmentation for robust visual question answering. In *ECCV*, 2022. 3

[6] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. In *NeurIPS*, 2021. 3

[7] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3

[8] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. *NeurIPS*, 2022. 3

[9] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[10] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *ECCV*, 2022. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5

[13] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPR*, 2020. 3

[14] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018. 5

[15] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020. 3

[16] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, 2022. 1, 2, 3, 4, 6, 7

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2, 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 6

[19] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, 2020. 3

[20] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *NeurIPS*, 2019. 3

[21] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, 2023. 1, 2, 3, 4, 6, 7

[22] 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*, 2019. 3, 6

[23] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 3, 6

[24] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, 2020. 3

[25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[26] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022. 3

[27] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *CVPR*, 2023. 3

[28] Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. Accurate and fast compressed video captioning. In *ICCV*, 2023. 3

[29] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, 2021. 1, 2, 3, 4, 6, 7

[30] 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*, 2021. 7

[31] 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 TCSVT*, 32(12):8238–8249, 2021. 1, 2, 6, 7

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 4

[33] Lan Wang, Gaurav Mittal, Sandra Sajeev, Ye Yu, Matthew Hall, Vishnu Naresh Boddeti, and Mei Chen. Protege: Untrimmed pretraining for video temporal grounding by video temporal grounding. In *CVPR*, 2023. 3

[34] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, 2022. 7

[35] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7

[36] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 3

[37] 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. 7

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 3

[39] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. Text-visual prompting for efficient 2d temporal video grounding. In *CVPR*, 2023. 3

[40] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. In *IJCAI*, 2020. 1, 2, 7

[41] 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 *CVPR*, 2020. 1, 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 *CVPR*, 2020. 2, 7

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1

[44] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 3

[45] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *CVPR*, 2023. 3