Towards Long-Form Spatio-Temporal Video Grounding

Xin Gu[1], Bing Fan[2], Jiali Yao[5], Zhipeng Zhang[3], Yan Huang[2], Cheng Han[4]
Heng Fan[2†], and Libo Zhang[5†]

[1]University of Chinese Academy of Sciences [2]UNT [3]Shanghai Jiao Tong University
[4]UMKC [5]Institute of Software Chinese Academy of Sciences

Abstract. In real scenarios, the videos can span several minutes or even hours, yet existing research on spatio-temporal video grounding (STVG), given a textual query, mainly focuses on localizing the target from a video of tens of seconds, typically *less than* one minute, hindering its applications. In this paper, we explore **L**ong-**F**orm **STVG** (**LF-STVG**), which aims to locate the target from long-term videos. In LF-STVG, long-term videos encompass a much longer temporal span and more irrelevant information, making it challenging for current short-form STVG approaches that process all the frames at once. Addressing these, we propose a novel **A**uto**R**egressive **T**ransformer architecture for LF-**STVG**, dubbed **ART-STVG**. Unlike current STVG methods requiring seeing the entire video sequence to make a full prediction at once, our ART-STVG regards the video as a streaming input and processes its frames sequentially, making it capable of easily handling the long videos. To capture spatio-temporal context in ART-STVG, spatial and temporal memory banks are developed and applied to decoders of ART-STVG. Considering that memories at different moments are not always relevant for localizing the target in current frame, we introduce simple yet effective memory selective strategies that enable the more relevant information for decoders, greatly improving the performance. Moreover, rather than parallelizing spatial and temporal localization as done in existing approaches, we introduce a novel cascaded spatio-temporal design that connects spatial decoder to temporal decoder during grounding, which allows ART-STVG to leverage more fine-grained target information to assist with complicated temporal localization in complex long videos, further boosting performance. On the newly extended datasets for LF-STVG, ART-STVG largely outperforms current approaches, while showing competitive results on Short-Form STVG. Our code is at: xxx.

Keywords: STVG · Long-Form STVG · Video Understanding

1 Introduction

Spatio-temporal video grounding (**STVG**) aims to localize a target of interest in *space* and *time* from an untrimmed video given a *free-form* textual query [54].

†Equal advising and corresponding authors

(a) Existing (SF-)STVG models: ***processing all frames in one time*** (b) Our LF-STVG model: ***processing one frame at a time***

Fig. 1: Comparison of existing STVG approaches (*e.g.*, [10, 12, 19, 26, 46, 50]) that see the entire video to make a full prediction at once in (a) and ART-STVG that processes frames one at a time and is more suitable for handling long videos in LF-STVG in (b).

As a multimodal task, it needs to accurately comprehend the spatio-temporal content of a video and make connections with the given textual query for target localization. Owing to its pivotal role in multimodal video understanding, STVG has recently attracted extensive attention (*e.g.*, [10,12,19,26,37,40,46,50,53,54]).

Despite advancements, current research mainly focuses on locating the target within a short-term video lasting tens of seconds, typically *less than* one minute. For instance, the average video length of existing popular benchmarks HCSTVG-v1/-v2 [40] and VidSTG [54] is 20 and 35 seconds, respectively. Nevertheless, in real-world applications, such as video retrieval and visual surveillance, the videos can span several *minutes* or even *hours*, resulting in a large *gap* between current research (focusing on target localization within *short-term* videos) and practical applications (the need of target localization in *long-term* videos). To mitigate this gap, in this paper, we explore **L**ong-**F**orm **STVG** (**LF-STVG**), which localizes the target of interest from *long-term* videos given a textual query.

To localize the desired target, existing STVG approaches (*e.g.*, [10, 12, 19, 26, 46, 50]) process *all* video frames in one time (see Fig. 1 (a)), aiming to capture and leverage global context of the entire video for localization. These approaches have achieved impressive results on current Short-Form STVG (SF-STVG). Nonetheless, as the video length increases, new challenges arise, prompting an **important question**: *Is processing all video frames at once, as done in existing SF-STVG models, still applicable to LF-STVG?*

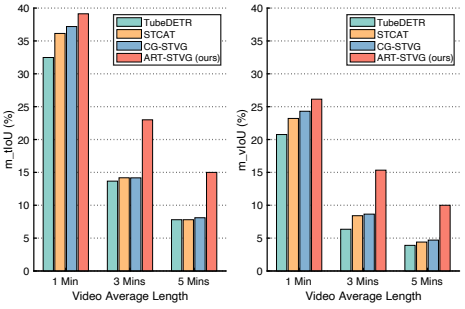


Fig. 2: Comparison on the datasets with minute-level videos. Existing STVG methods severely suffer localization in long sequences, and ART-STVG significantly surpasses these models. Moreover, the longer the video is, the larger the improvement of ART-STVG over other methods is.

Our response is **negative**! In LF-STVG, the long video often comprises a longer temporal span, which largely increases the complexities of spatio-temporal localization. In addition, the long video commonly contains far more irrelevant information, requiring a model to identify the target event from extensive redundant content. For these reasons,

processing all frames of a long video at once, as done in current STVG methods, presents significant challenges in capturing long-term spatio-temporal relationships and handling excessive irrelevant information for accurate localization (see Fig. 2). Additionally, it causes computational bottlenecks because of high GPU memory requirements for simultaneous target localization in all frames.

Addressing the aforementioned challenges, we propose a novel **A**uto**R**egressive **T**ransformer method for LF-**STVG**, dubbed **ART-STVG**. Specifically, it treats the video as a streaming input, and processes its frames sequentially (see Fig. 1 (b)). To capture the crucial spatio-temporal contextual information in the video, we maintain two memory banks, that reserve essential spatio-temporal information from the video, for spatial and temporal decoders in ART-STVG. Since the memories in the banks are *not* equally important to a certain frame, we introduce simple yet effective memory selective strategies to leverage more relevant information in the memory banks for grounding, effectively boosting performance. Compared to existing approaches that require seeing the entire video for prediction, our proposed ART-STVG ingests frames one at a time for prediction, hence naturally processing longer videos and resolving the computational bottleneck faced by current approaches. Furthermore, rather than parallelizing the spatial and temporal localization as is done in existing approaches, we propose a novel cascaded spatio-temporal design which connects spatial decoder to temporal decoder during grounding. By doing this, our ART-STVG can enjoy more fine-grained target information from the spatial decoder to assist with the more complicated temporal localization, further enhancing performance. To our best knowledge, this paper is the *first* to explore the LF-STVG problem, and ART-STVG is the *first* framework attempting to handle LF-STVG.

To verify the effectiveness of our ART-STVG in long-term videos, we extend the validation set of the short-term benchmark HCSTVG-v2 [40] (the reason for choosing HCSTVG-v2 for the extension is described later). Specifically, we extend its average video length from 20 seconds to 1/3/5 minutes, hence it is referred to as LF-STVG-1min/3min/5min. We conduct extensive experiments on both long-form and short-form STVG. The results demonstrate that our ART-STVG outperforms all existing approaches on LF-STVG by achieving new state-of-the-art results while showing competitive performance on SF-STVG.

In summary, our **contributions** are as follows: ♠ We explore the LF-STVG problem and introduce a novel memory-augmented autoregressive transformer, dubbed ART-STVG, for LF-STVG; ♥ We propose memory selection strategies that allow the selection of relevant crucial spatio-temporal context to enhance target localization; ♣ We introduce a cascaded spatio-temporal decoder design to fully leverage the fine-grained information produced by spatial localization to assist in temporal localization; ♦ In extensive experiments on both long-term and short-term benchmarks, our ART-STVG achieves excellent performance.

2 Related Work

Spatio-temporal video grounding aims to locate the tube in the video that corresponds to a given textual query. Early methods (*e.g.*, [37,39,45,51,54]) are

predominantly two-stage approaches. These approaches first use a pre-trained object detector [34] to generate object proposals and then select the proposals based on the given textual query. Such methods are easily limited by the pre-trained object detector. Recent methods (*e.g.*, [10, 12, 19, 26, 38]), inspired by DETR [2], propose one-stage frameworks that directly generate tubes for target localization, performing better than two-stage models. Nevertheless, both early two-stage and recent one-stage approaches focus on SF-STVG and process the entire video at one time for simultaneous target localization in all frames. ***Different from*** existing approaches, our ART-STVG is specially designed for LF-STVG. Specifically, ART-STVG treats the video as a streaming input and processes its frames sequentially with an autoregressive framework, thus making it more suitable for handling long-term videos. Besides, we design memory selection strategies specifically for this streaming framework to enhance localization.

Long-term video understanding has received significant attention recently in various tasks such as action detection (*e.g.*, [3,42]), video captioning (*e.g.*, [18, 24]), and video question answering (*e.g.*, [4, 13, 30, 36]). The major challenge of long video understanding is that capturing complex spatio-temporal dependencies over long durations requires a high computational cost. To address this, early methods (*e.g.*, [6, 8, 17, 47]) model pre-extracted video features without jointly training the backbone. Recent works (*e.g.*, [1, 49, 52]) present efficient strategies to process more frames simultaneously, while others (*e.g.*, [7, 13, 33, 44, 48]) construct streamlined transformers and integrate memory banks for effective long video understanding. ***Different from*** these works, we focus on LF-STVG. Particularly, ***unlike*** video query answering [13, 36] in which the memory is global context information for the entire video, the memory in our ART-STVG is text-guided spatial instance and temporal event boundary cues, specifically designed for LF-STVG. In addition, for memory selection, our method merges spatial memories using text as guidance and temporal memories using event boundary cues from videos, while the other methods [13,36] do not have these mechanisms because they aim at different tasks.

Autoregressive architecture has been extensively studied and applied in various domains. Early autoregressive models include recurrent neural networks (RNN) [31], long-short-term memory networks (LSTM) [9, 16], and gated recurrent units (GRU) [5]. Recently, autoregressive transformer models (*e.g.*, [22, 25, 27, 35, 41, 43]) by applying self-attention mechanism have further advanced the field by enabling serial computation and capturing the long-range dependencies. ***Different from*** the aforementioned methods, in this paper we propose an autoregressive transformer architecture for LF-STVG.

3 The Proposed Approach

Overview. We propose ART-STVG, a memory-augmented autoregressive transformer for LF-STVG. As shown in Fig. 3, the framework begins with a multi-modal encoder (Sec. 3.1) that extracts and fuses visual and textual features. Following this, the cascaded spatio-temporal decoder performs autoregressive

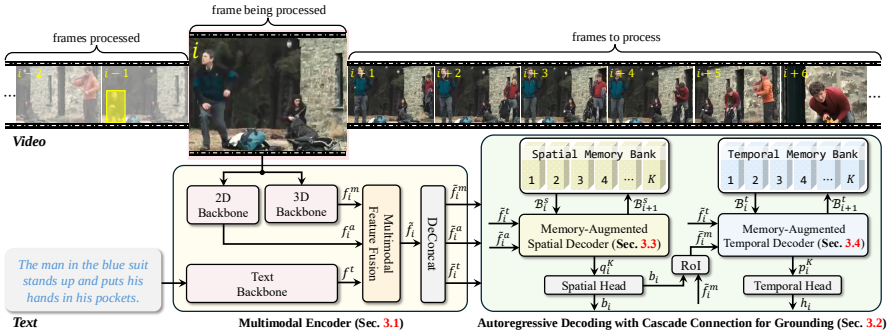


Fig. 3: Overview of ART-STVG, comprising a multimodal encoder and an autoregressive decoder with cascade connection for target localization frame by frame.

decoding for grounding (Sec. 3.2). Specifically, the memory-augmented spatial decoder (Sec. 3.3) captures the spatial location information of the target, while the memory-augmented temporal decoder (Sec. 3.4) focuses on learning the temporal location information.

Since ART-STVG processes frames sequentially, in the following, we take the processing of the i^{th} frame as an example to illustrate our ART-STVG.

3.1 Multimodal Encoder

Given the i^{th} frame of a video and the corresponding textual query, the multimodal encoder generates a multimodal feature, which is sent to the autoregressive decoder for localization. The multimodal encoder comprises feature extraction and feature fusion, as described below.

Feature Extraction. For the i^{th} frame of the video sequence, we separately extract its 2D appearance and 3D motion features to leverage rich static and dynamic cues. Specifically, the appearance feature is extracted using ResNet-101 [14], and the motion feature is extracted through VidSwin [29]. Please ***note***, when applying VidSwin to extract motion features, the $(i-1)^{\text{th}}$ frame is also used as input. The appearance feature of frame i is denoted as $f_i^a \in \mathbb{R}^{H \times W \times C_a}$, where H, W, and C_a are height, width, and channel dimensions. Similarly, the motion feature is represented as $f_i^m \in \mathbb{R}^{H \times W \times C_m}$ with C_m the channel dimension. For the text, we first tokenize it to a word sequence, and then apply RoBERTa [28] to extract its feature $f^t \in \mathbb{R}^{N_t \times C_t}$, where N_t is the text feature length and C_t represents the channel dimension.

Feature Fusion. Different modalities typically contain complementary information. Therefore, we fuse the appearance feature f_i^a and motion feature f_i^m of the i^{th} video frame with the textual feature f^t to generate a multimodal feature of the i^{th} frame. Specifically, we first project them to the same channel dimension C, and then concatenate them to produce the multimodal feature $f_i^{'}$, as follows,

$$f_i^{'} = [\underbrace{f_{i_1}^a, f_{i_2}^a, ..., f_{i_{H\times W}}^a}_{\text{appearance feature } f_i^a}, \underbrace{f_{i_1}^m, f_{i_2}^m, ..., f_{i_{H\times W}}^m}_{\text{motion feature } f_i^m}, \underbrace{f_1^t, f_2^t, ..., f_{N_t}^t}_{\text{textual feature } f^t}]$$

Afterwards, we apply a self-attention encoder [43] to further fuse the multimodal features, as follows,

$$\tilde{f}_i = \texttt{SelfAttEncoder}(f_i^{'} + \mathcal{E}_{pos} + \mathcal{E}_{typ})$$

where $\mathcal{E}_{pos}$ and $\mathcal{E}_{typ}$ denote position and type embeddings, and $\texttt{SelfAttEncoder}(\cdot)$ is the self-attention encoder [43] with N (N=6) self-attention encoder blocks.

After obtaining the $\tilde{f}_i$, we deconcatenate it to generate enhanced appearance, motion, and textual features $\tilde{f}_i^a$, $\tilde{f}_i^m$, and $\tilde{f}_i^t$ via $[\tilde{f}_i^a, \tilde{f}_i^m, \tilde{f}_i^t] = \texttt{DeConcat}(\tilde{f}_i)$ and apply them in the subsequent decoder for target localization.

3.2 Autoregressive Decoding for Grounding

ART-STVG autoregressively decodes video frames to sequentially predict spatio-temporal positions of target. As displayed in Fig. 3, the decoding process of ART-STVG contains two parts, including spatial grounding and temporal grounding via two decoders. The former is responsible for predicting the spatial location of the target, while the latter generates the temporal location of the target event. To capture spatio-temporal context in ART-STVG, spatial and temporal memory banks storing historical information, with effective memory selection strategies, are developed and applied in the grounding process, largely enhancing performance. Besides, *rather than paralleling* the spatial and temporal grounding as done in current approaches, we propose a novel ***cascaded*** design to connect spatial and temporal grounding in ART-STVG (see decoding part in Fig. 3). Such cascaded spatio-temporal architecture allows ART-STVG to employ more fine-grained target cues, provided by spatial grounding, to assist with temporal localization in complex long videos, further improving ART-STVG for LF-STVG.

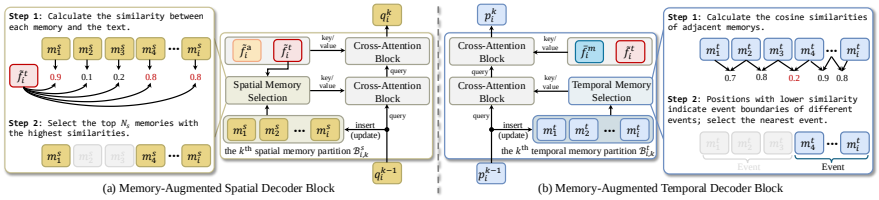
Spatial Grounding. In ART-STVG, the spatial grounding is achieved by learning a spatial query via iterative interaction with the multimodal feature. Let q_i^0 denote the initial spatial query in the i^{th} frame and $\mathcal{B}_i^s$ is the spatial memory bank at this moment. Given the appearance feature $\tilde{f}_i^a$ and textual feature $\tilde{f}_i^t$ from $\tilde{f}_i$ in frame i, the interaction of the spatial query with multimodal feature is achieved as follows,

$$q_i^K, \mathcal{B}_{i+1}^s = \texttt{MA-SpatialDecoder}(q_i^0, \mathcal{B}_i^s, [\tilde{f}_i^a, \tilde{f}_i^t])$$

where $\texttt{MA-SpatialDecoder}(\cdot)$ is the memory-augmented spatial decoder with K spatial decoder blocks (described in Sec. 3.3). It is worth ***noting*** that, the spatial memory bank $\mathcal{B}_i^s$ contains K (*i.e.*, the number of decoder blocks) partitions, with each partition corresponding to a spatial decoder block. q_i^K represents the final spatial query feature after K decoder blocks, and $\mathcal{B}_{i+1}^s$ the new memory bank updated with spatial information from frame i (see Sec. 3.3). After this, a spatial head, containing an MLP module, is used to predict the final object box b_i, as follows,

$$b_i = \texttt{SpatialHead}(q_i^K)$$

where $b_i \in \mathbb{R}^4$ represents the central position, width, and height of the predicted target box in the i^{th} frame.



(a) Memory-Augmented Spatial Decoder Block (b) Memory-Augmented Temporal Decoder Block

Fig. 4: Illustration of the architectures for memory-augmented spatial decoder block in (a) and temporal decoder block in (b).

Temporal Grounding. Similar to spatial grounding, we learn a temporal query for temporal grounding by interacting with the multimodal feature. To exploit fine-grained spatial target information to assist with temporal grounding, we design a cascade architecture. Specifically, with target box b_i from spatial grounding, we first extract fine-grained target motion feature $\bar{f}_i^m \in \mathbb{R}^{1\times1\times\mathcal{C}}$ using RoI pooling [34], as follows,

$$\bar{f}_i^m = \texttt{RoI}(\tilde{f}_i^m, b_i)$$

Compared to $\tilde{f}_i^m$, $\bar{f}_i^m$ focuses more on target region and thus benefits localization.

After this, we interact the temporal query with multimodal feature. Let p_i^0 denote the initial temporal query in frame i and $\mathcal{B}_i^t$ the temporal memory bank at this moment. With fine-grained motion feature $\bar{f}_i^m$ and textual feature $\tilde{f}_i^t$, the interaction of temporal query and multimodal feature is performed as follows,

$$p_i^K, \mathcal{B}_{i+1}^t = \texttt{MA-TemporalDecoder}(p_i^0, \mathcal{B}_i^t, [\bar{f}_i^m, \tilde{f}_i^t])$$

where $\texttt{MA-TemporalDecoder}(\cdot)$ is the memory-augmented temporal decoder with K temporal decoder blocks (described in Sec. 3.4). Similar to $\mathcal{B}_i^s$, the temporal memory bank $\mathcal{B}_i^t$ also comprises K partitions, with each corresponding to a temporal decoder block. p_i^K is the final temporal query feature after the decoder, and $\mathcal{B}_{i+1}^t$ the new memory bank updated with temporal information in frame i (see Sec. 3.4). After this, a temporal head implemented with an MLP module is adopted for temporal localization in frame i, as follows,

$$h_i = \texttt{TemporalHead}(p_i^K)$$

where $h_i \in \mathbb{R}^2$ contains event start and end probabilities h_i^s and h_i^e for frame i.

By sequentially performing spatial and temporal grounding, we achieve target localization in each frame i, and meanwhile adopt information in frame i to update the memory banks for target localization in frame $(i{+}1)$.

3.3 Memory-Augmented Spatial Decoder

We propose a memory-augmented spatial decoder, guided by spatial memory from the spatial memory bank, to learn the target spatial position from the multimodal feature. Specifically, the memory-augmented spatial decoder comprises K decoder blocks in a cascade for spatial grounding. As shown in Fig. 4 (a), each spatial decoder block corresponds to a partition in the spatial memory and contains two cross-attention blocks [43]. Concretely, in the k^{th} $(1 \le k \le K)$

Fig. 5: Comparison of attention maps of the spatial query *without* and *with* using selective spatial memory. The red box indicates the target object to be located. Through comparison, we observe the use of selective spatial memory helps the model focus more on the target regions, which benefits the final target localization.

spatial decoder block, given the appearance feature $\tilde{f}_i^a$, the textual feature $\tilde{f}_i^t$, and the spatial query q_i^{k-1} (q_i^0 initialized by zeros) of the i^{th} frame, we first perform memory selection and then apply the selected memory to enhance the spatial query feature in spatial decoding.

Spatial Memory Selection. Since the spatial query contains crucial target information, we first insert the spatial query q_i^{k-1} into the k^{th} partition of spatial memory bank $\mathcal{B}_{i,k}^s$ corresponding to the k^{th} decoder block. Please ***note*** that, this insertion procedure also completes the update of each partition $\mathcal{B}_{i,k}^s$ in $\mathcal{B}_i^s$ to $\mathcal{B}_{i+1,k}^s$ in $\mathcal{B}_{i+1}^s$. Or in other words, we update the memory bank by simply adding the query as a new memory, without removing any existing memories.

After this, we perform memory selection from $\mathcal{B}_{i+1,k}^s$ for decoder block k. The *motivation* behind this selection is, the memories at different moments are not always relevant for target localization in current frame, and selecting more relevant information in the spatial decoding enables learning better query feature for grounding. Specifically, the selective spatial memory $\mathcal{M}_{i,k}^s$ for block k can be obtained via two steps in memory selection: *first*, we calculate the similarity between each spatial memory and the textual feature; *second* , based on the similarity scores, the top N_s spatial memories with the highest scores are selected to form $\mathcal{M}_{i,k}^s$. Fig. 4 (a)-left illustrates this spatial memory selection process.

Memory-Augmented Spatial Decoding. During decoding, we send q_i^{k-1} to decoder block k for learning q_i^k. To exploit spatial context, we first interact the query with selective spatial memory through a cross-attention block, as follows,

$$\tilde{q}_i^{k-1} = \texttt{CrossAtt}(q_i^{k-1}, \mathcal{M}_{i,k}^s)$$

where $\tilde{q}_i^{k-1}$ denotes the memory-augmented query feature in decoder block k, and $\texttt{CrossAtt}(\mathbf{u}, \mathbf{v})$ is the cross-attention block [43], with $\mathbf{u}$ generating query and $\mathbf{v}$ key/value. After this, we further interact $\tilde{q}_i^{k-1}$ with the multimodal appearance and textual features for learning q_i^k, as follows,

$$q_i^k = \texttt{CrossAtt}(\tilde{q}_i^{k-1}, [\tilde{f}_i^a, \tilde{f}_i^t])$$

where q_i^k is the learned query feature, which is sent to the next decoder block for further query feature learning.

In Fig. 5, we compare the attention map of the spatial query with and without using selective spatial memory. We can see using selective spatial memory helps

Fig. 6: Illustration of selective temporal memory. In the middle figure, the attention sequence indicates cosine similarities of adjacent memories. The lower similarity (greener color) indicates the potential event boundaries. Besides, in the up figure, we show predicted start and end probabilities, which are accurate to capture the target event. The red box denotes the ground truth corresponding to the text query.

the model focus more on target regions for better localization. After K spatial decoder blocks, the spatial query feature q_i^K is generated for spatial prediction.

3.4 Memory-Augmented Temporal Decoder

The memory-augmented temporal decoder learns target temporal position using temporal memory from temporal memory bank. It has K blocks for temporal grounding, with each corresponding to a temporal memory partition and containing two cross-attention blocks, as in Fig. 4 (b). In temporal decoder block k ($1 \leq k \leq K$), given motion and textual features $\bar{f}_i^m$ and $\tilde{f}_i^t$, and temporal query p_i^{k-1} (p_i^0 initialized by zeros), we first perform temporal memory selection and then apply the selected memory to enhance temporal decoding.

Temporal Memory Selection. The temporal query p_i^{k-1} contains temporal event information. Thus, we first insert it into the k^{th} partition of temporal memory bank $\mathcal{B}_{i,k}^t$ (also updating $\mathcal{B}_{i,k}^t$ to $\mathcal{B}_{i+1,k}^t$). Since long-term videos often contain multiple events, selecting relevant temporal memory related to the current event helps the temporal decoding better locate the event boundaries. To achieve this and obtain selective temporal memory $\mathcal{M}_{i,k}^t$, inspired by TextTiling [15], we perform two steps in temporal memory section: in the *first* step, we calculate the similarities between the memories of adjacent frames; in the *second* step, points with lower similarities are considered the event boundaries between different events, and we only select the memories corresponding to the event closest to the current frame, as shown in Fig. 4 (b)-right.

Memory-Augmented Temporal Decoding. In decoding, we send the temporal query p_i^{k-1} to temporal decoder block k for learning p_i^k. To exploit temporal context for enhancing query learning, we first interact the query with the selective temporal memory $\mathcal{M}_{i,k}^t$ by a cross attention block as follows,

$$\tilde{p}_i^{k-1} = \texttt{CrossAtt}(p_i^{k-1}, \mathcal{M}_{i,k}^t)$$

where $\tilde{p}_i^{k-1}$ represents the memory-augmented query feature in decoder block k. After this, we further interact $\tilde{p}_i^{k-1}$ with multimodal motion and textual features, as follows,

$$p_i^k = \texttt{CrossAtt}(\tilde{p}_i^{k-1}, [\bar{f}_i^m, \tilde{f}_i^t])$$

where p_i^k represents the learned query feature, which is sent to the next decoder block for further query feature learning. As shown in Fig. 6, the temporal memory selection strategy segments the video into different events and selects the memory closest to current moment, benefiting localization of target event. After K blocks in decoder, the temporal query feature p_i^K is adopted for temporal prediction.

3.5 Optimization

Given the video containing N_f frames and its textual query, ART-STVG predicts (1) the object boxes $\mathcal{B} = \{b_i\}_{i=1}^{N_f}$ in the memory-augmented spatial decoder and event start timestamps $\mathcal{H}_s = \{h_i^s\}_{i=1}^{N_f}$ and end timestamps $\mathcal{H}_e = \{h_i^e\}_{i=1}^{N_f}$ in the memory-augmented temporal decoder. During training, with the groundtruth of the bounding box $\mathcal{B}^*$, start timestamps $\mathcal{H}_s^*$ and end timestamps $\mathcal{H}_e^*$, we can calculate the total loss $\mathcal{L}$ af follows,

$$\mathcal{L} = \underbrace{\lambda_k(\mathcal{L}_{\text{KL}}(\mathcal{H}_s^*, \mathcal{H}_s) + \mathcal{L}_{\text{KL}}(\mathcal{H}_e^*, \mathcal{H}_e))}_{\text{loss of memory-augmented temporal decoder}} + \underbrace{\lambda_l \mathcal{L}_1(\mathcal{B}^*, \mathcal{B}) + \lambda_u \mathcal{L}_{\text{IoU}}(\mathcal{B}^*, \mathcal{B})}_{\text{loss of memory-augmented spatial decoder}}$$

where $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_1$ and $\mathcal{L}_{\text{IoU}}$ are KL divergence, smooth L1 and IoU losses. λ_k, λ_l and λ_u are parameters to balance the training loss.

4 Experiments

Implementation. Our ART-STVG is implemented based on PyTorch [32]. We use ResNet-101 [14] for appearance feature extraction, VidSwin-tiny [29] for motion feature extraction, and RoBERTa-base [28] for textual feature extraction. Following previous work [10, 19, 26], we use pre-trained MDETR [21] to initialize the appearance and text backbones and the multimodal fusion module. The hidden dimension of the encoder and decoder is $C = 256$, with channel dimensions of $C_a = 2048$ for appearance features, $C_m = 768$ for motion features, and $C_t = 768$ for textual features. We sample video frames at FPS of 3.2 and resize each frame to have a short side of 420. The video frame length during training is $N_f = 64$, and the text sequence length is $N_t = 30$. During training, we adopt Adam [23] with an initial learning rate of $1e-5$ for the pre-trained backbone and $1e-4$ for other modules, while keeping the motion backbone frozen. Similar to [10, 19, 26], the parameters λ_k, λ_l and λ_u are empirically set to 10, 5, and 3.

Datasets. Since there are no available benchmarks available for LF-STVG, we opt to extend HCSTVG-v2 [40] for creating new datasets for LF-STVG. The ***reason*** for choosing HCSTVG-v2 only for extension is that it is the ***only*** benchmark which provides available source videos, thus allowing for extension with longer videos. Specifically, HCSTVG-v2 originally contains 16,000 video-sentence pairs in complex scenes, including 10,131 training samples, 2,000 validation samples, and 4,413 testing samples. Each video lasts 20 seconds and is paired with a natural query sentence averaging 17.25 words. As annotations of the test set are not publicly available, the results are reported on the validation set, as in existing

Table 1: Comparison to other methods on long-term videos. The best result is highlighted in the **bold** font.

	LF-STVG-1min				LF-STVG-3min				LF-STVG-5min			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
TubeDETR [50]	32.5	20.8	25.7	8.7	13.6	6.4	7.2	2.9	7.8	3.9	0.7	0.1
STCAT [19]	36.1	23.2	34.4	10.4	14.2	8.4	3.0	0.1	7.8	4.4	0.3	0.0
CG-STVG [11]	37.2	24.3	32.6	10.9	14.2	8.7	3.2	0.3	8.1	4.7	0.3	0.0
TA-STVG [12]	38.4	25.2	35.5	12.1	13.9	8.5	3.3	0.2	7.7	4.5	0.3	0.0
Baseline (ours)	30.1	19.7	25.5	8.3	16.2	10.7	10.5	4.5	9.2	5.3	4.5	1.1
ART-STVG (ours)	39.1	26.1	36.8	17.6	23.0	15.3	20.1	9.5	15.0	10.0	11.4	4.7
Δ over Baseline	+9.0	+6.4	+11.3	+9.3	+6.8	+4.6	+9.6	+5.0	+5.8	+4.7	+6.9	+3.6

methods [10,26,50]. For this reason, we extend only the validation set to lengths of 1/3/5 minutes, referred to as LF-STVG-1min/3min/5min, for the evaluation of LF-STVG. The extensions are based on original YouTube videos, not concatenated clips and we extend the HCSTVG-v2 val clips by randomly padding frames on either side. We only constrain total video length, while randomizing extension duration and direction. Three domain-related students manually verify that each grounding tube remains unique and aligned with the textual query. The split and evaluation protocol follow the HCSTVG-v2. It is worth ***noting*** that we do not change the training set, and ART-STVG is trained using the same samples as in short-form STVG approaches for a fair comparison.

Metrics. Follow previous works [10,19,26], we utilize the commonly used metrics m_tIoU, m_vIoU, and vIoU@R for evaluation. m_tIoU evaluates the effectiveness of temporal grounding by averaging tIoU scores across the entire test set. m_vIoU assesses spatial grounding performance by averaging vIoU scores. Additionally, vIoU@R measures performance by determining the proportion of test samples with vIoU scores exceeding a threshold R. For a more detailed explanation of the evaluation metrics used, please refer to previous works [10, 19, 26].

4.1 Comparison on Long-Form STVG

We compare ART-STVG with other methods on the extended LF-STVG benchmarks with different average video lengths. Please ***note*** that all methods including our ART-STVG are trained *exclusively* on the HCSTVG-v2 training set (*i.e.*, the average length of training videos is 20 seconds) for fair comparison.

Tab. 1 reports the results. As shown in Tab. 1, our ART-STVG significantly outperforms existing STVG approaches in all metrics on all three benchmarks, showing the superiority of our method in grounding target in long-term videos. Specifically, ART-STVG outperforms TA-STVG by achieving improvements in m_tIoU/m_vIoU scores with 0.7%/0.9%, 9.1%/6.8%, and 7.3%/5.5% gains on three different video lengths, respectively. In addition, compared to our baseline, which has a similar architecture to ART-STVG but *without* memory and memory selection modules (please kindly check its architecture in ***supplementary***

Table 2: Ablation studies of the selective temporal memory in the temporal decoder. TM: Temporal Memory; TMS: Temporal Memory Selection.

	TM	TMS	m_tIoU	m_vIoT	vIoU@0.3	vIoU@0.5
❶	-	-	16.7	11.1	11.9	4.7
❷	✓	-	9.6	6.2	4.7	1.5
❸	✓	✓	**23.0**	**15.3**	**20.1**	**9.5**

Table 3: Ablation studies of the selective spatial memory in the spatial decoder. SM: Spatial Memory; SMS: Spatial Memory Selection.

	SM	SMS	m_tIoU	m_vIoT	vIoU@0.3	vIoU@0.5
❶	-	-	21.3	13.9	16.4	8.0
❷	✓	-	22.1	14.2	17.0	9.0
❸	✓	✓	**23.0**	**15.3**	**20.1**	**9.5**

Table 4: Ablation studies of different designs for decoders in ART-STVG.

	Design	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	Parallel	21.5	13.9	17.3	8.2
❷	Cascaded	**23.0**	**15.3**	**20.1**	**9.5**

Table 5: Ablation studies on N_s.

		m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	$N_s = 16$	22.7	15.0	18.4	9.2
❷	$N_s = 32$	**23.0**	**15.3**	**20.1**	**9.5**
❸	$N_s = 48$	22.5	14.7	18.2	9.1

material), ART-STVG achieves remarkable improvements in m_tIoU/m_vIoU with 9.0%/6.4%, 6.8%/4.6%, and 5.8%/4.7% gains on the three benchmarks, evidencing the importance of selective memories for LF-STVG.

4.2 Ablation Study

To better understand ART-STVG, we further study its different designs by conducting ablation experiments on LF-STVG-3min with in-depth analysis. Our final configuration is highlighted in cyan.

Impact of selective temporal memory. We set up a temporal memory bank in the temporal decoder to store temporal contextual information of target event and use this temporal memory to assist in locating the start and end of event related to the target. To verify its effectiveness for LF-STVG, we conduct an ablation study in Tab. 2. As in Tab. 2, without temporal memory, our method achieves a m_tIoU score of 16.7% (❶). When incorporating all temporal memories from the memory bank, the m_tIoU score is decrease to 9.6% (❶ *v.s.* ❷). This is because the long-term video often contains multiple events, and using all temporal memories can introduce a lot of irrelevant information. With the help of memory selection, the m_tIoU score is improved to 23.0% with 13.4% gains (❷ *v.s.* ❸). These experimental results show that our selective temporal memory can effectively enhance ART-STVG for improving LF-STVG in long videos.

Impact of selective spatial memory. Similar to the temporal decoder, we adopt a spatial memory bank in the spatial decoder to learn contextual information about the target for locating its spatial position. We conduct an ablation in Tab. 3. From Tab. 3, we observe that integrating all spatial memories can improve the m_tIoU score to 22.1% with 0.8% gains (❶ *v.s.* ❷), and applying the memory selection strategy can further enhance the m_tIoU score to 23.0% with 0.9% gains (❷ *v.s.* ❸), validating the importance of selective memory.

Impact of design for spatial and temporal decoders. Unlike existing STVG approaches that parallelize spatial and temporal decoders, we introduce

a novel cascaded spatio-temporal design for LF-STVG. A key challenge of LF-STVG is that, as video length grows, the difficulty of temporal localization increases. To handle this, we cascade spatial and temporal decoders in ART-STVG, which allows the use of spatial localization information to assist temporal localization. We conduct an ablation in Tab. 4. We can see that, cascading spatial and temporal decoders outperforms the parallel design, with improvements of 1.5% and 1.4% scores on m_tIou and m_vIoU (❶ *v.s.* ❷), respectively.

Impact of the number of selective spatial memories. In the spatial decoder, we utilize N_s to control the number of selective spatial memories. To explore the impact of N_s, we conducted the ablation experiment in Tab. 5. We can see that when N_s is 32, the performance of the model is the best (❷).

Impact of training video length. To investigate the impact of training videos of different lengths, we extend HCSTVG-v2 training set to 40 seconds and use it to train both existing methods and ART-STVG. As shown in Tab. 6, we can observe that all methods show improvements when trained on 40-second videos compared to 20-second videos (Tab. 6 *v.s.* Tab. 1 on LF-STVG-3min). This shows that training with longer videos enhances target localization in long-term videos, yet results in significantly increasing training costs. More importantly, our method still achieves the best performance on all metrics, with improvements of 7.6% score on m_tIoU and 7.0% score on m_vIoU compared to TA-STVG.

Table 6: Comparison of ART-STVG with existing methods on LF-STVG using the HCSTVG-v2 validate set. The best result is highlighted in the **bold** font.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
TubeDETR [50]	20.8	11.5	9.8	3.9
STCAT [19]	21.0	12.2	7.4	0.6
CG-STVG [10]	20.5	12.0	8.0	1.0
TA-STVG [12]	20.7	11.8	7.7	0.5
ART-STVG (ours)	**28.3**	**18.8**	**27.0**	**11.9**

4.3 Comparison on Short-Form STVG

We further evaluate ART-STVG on SF-STVG in Tab. 7 on HCSTVG-v2 validation set. As in Tab. 7, our method shows competitive results to current STVG methods on short-term videos. Current methods use non-autoregressive structures that process video frames in parallel to capture inter-frame relationships, and are specially designed for target localization in short-term videos. Despite this, our ART-STVG, adopting an autoregressive structure, outperforms most existing methods, falling only behind the latest state-of-the-art TA-STVG [12] by 1.2% in m_tIoU and 1.0% in m_vIoU. Moreover, our method shows clear gains compared to the baseline without memory.

4.4 Qualitative Results

To qualitatively validate our method on LF-STVG, we present grounding results and comparisons with the baseline method without memory on the LF-STVG-3min benchmark. From Fig. 7, we observe that the baseline without memory often locates inconsistent targets across different video frames, such as in the

Table 7: Comparison of ART-STVG with existing methods on LF-STVG using the HCSTVG-v2 validate set. The best result is highlighted in the **bold** font.

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
2D-Tan [39]	-	30.4	50.4	18.8
MMN [45]	-	30.3	49.0	25.6
TubeDETR [50]	53.9	36.4	58.8	30.6
STCAT [19]	56.6	36.9	60.3	33.6
STVGFormer [26]	58.1	38.7	65.5	33.8
CG-STVG [10]	60.0	39.5	64.5	36.3
TA-STVG [12]	**60.4**	**40.2**	**65.8**	**36.7**
Baseline (ours)	46.2	29.9	45.0	21.1
ART-STVG (ours)	59.2	39.2	64.4	33.2

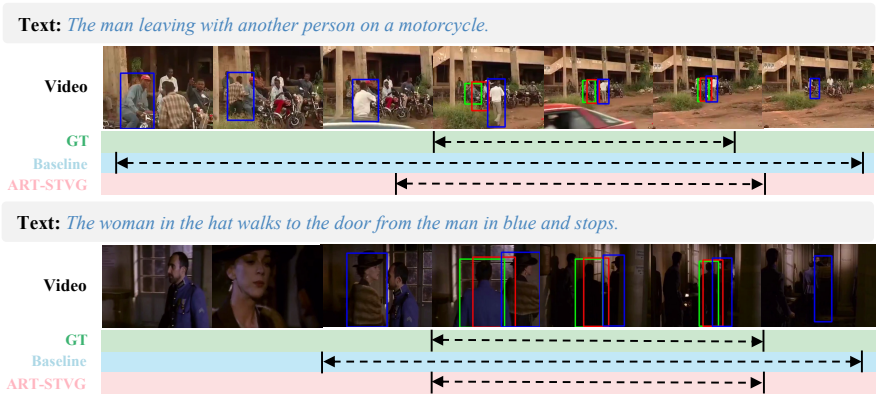


Fig. 7: Qualitative comparison of our ART-STVG and baseline.

third and fifth examples. In contrast, our method locates the target more consistently and accurately, demonstrating the effectiveness of our approach.

4.5 Analysis of Efficiency

To analyze the efficacy and complexity, we report the efficiency of our model and its comparison with other methods in Tab. 8. As shown in Tab. 8, our model size is similar to other methods. Although our inference time (for 64 images, aligned to SF-STVG methods) is longer due to autoregressive processing, at 1.09 seconds, compared to the inference times of STCAT, CG-STVG, and TA-STVG, which are 0.47, 0.71 and 0.69 seconds respectively, our GPU memory usage is much lower than other methods. Specifically, our GPU memory usage is 7.9G, while other methods such

Table 8: Comparison on the model efficacy and complexity on a single A100 GPU.

Methods	Params		Inference	
	Trainable	Total	Time	GPU Mem
TubeDETR [50]	185 M	185 M	0.23 s	22.1 G
STCAT [20]	207 M	207 M	0.47 s	23.6 G
CG-STVG [10]	203 M	231 M	0.71 s	25.9 G
TA-STVG [12]	206 M	234 M	0.69 s	25.1 G
ART-STVG (ours)	207 M	235 M	1.09 s	7.9 G

as TA-STVG and CG-STVG have GPU memory usages of 25.1G and 25.9G, respectively. Therefore, our method is more suitable for handling long videos.

Due to limited space, we show additional results, analyzes, and discussions in the ***supplementary material***.

5 Conclusion

In this work, we explore the LF-STVG problem, and propose ART-STVG to handle long-term videos. The crux of ART-STVG lies in the simple but effective use of selective memories, which are applied to decoders to leverage crucial spatio-temporal contextual information for target localization, greatly improving performance. Additionally, our proposed cascaded spatio-temporal decoder design effectively leverages spatial localization to assist temporal localization in long-term videos. To validate the effectiveness of ART-STVG, we conduct extensive experiments on LF-STVG benchmarks. The results show that ART-STVG significantly outperforms other methods. In addition, our method achieves comparable performance on SF-STVG benchmarks, demonstrating its generality. We hope that this work can inspire more future research on LF-STVG.

Acknowledgment. BF, YH, CH, and HF are not supported by any funds this this work.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv (2023) 4
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 4
3. Cheng, F., Bertasius, G.: Tallformer: Temporal action localization with a long-memory transformer. In: ECCV (2022) 4
4. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv (2024) 4
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv (2014) 4
6. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015) 4
7. Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR (2019) 4
8. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: CVPR (2017) 4
9. Graves, A., Graves, A.: Long short-term memory. Supervised sequence labelling with recurrent neural networks pp. 37–45 (2012) 4
10. Gu, X., Fan, H., Huang, Y., Luo, T., Zhang, L.: Context-guided spatio-temporal video grounding. In: CVPR (2024) 2, 4, 10, 11, 13, 14

11. Gu, X., Fan, H., Huang, Y., Luo, T., Zhang, L.: Context-guided spatio-temporal video grounding. In: CVPR (2024) 11
12. Gu, X., Shen, Y., Luo, C., Luo, T., Huang, Y., Lin, Y., Fan, H., Zhang, L.: Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. In: ICLR (2025) 2, 4, 11, 13, 14
13. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: CVPR (2024) 4
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 5, 10
15. Hearst, M.A.: Text tiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics **23**(1), 33–64 (1997) 9
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997) 4
17. Hussein, N., Gavves, E., Smeulders, A.W.: Timeception for complex action recognition. In: CVPR (2019) 4
18. Islam, M.M., Ho, N., Yang, X., Nagarajan, T., Torresani, L., Bertasius, G.: Video recap: Recursive captioning of hour-long videos. In: CVPR (2024) 4
19. Jin, Y., Li, Y., Yuan, Z., Mu, Y.: Embracing consistency: A one-stage approach for spatio-temporal video grounding. In: NeurIPS (2022) 2, 4, 10, 11, 13, 14
20. Jin, Y., Yuan, Z., Mu, Y., et al.: Embracing consistency: A one-stage approach for spatio-temporal video grounding. NeurIPS (2022) 14
21. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV (2021) 10
22. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: ICML (2020) 4
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 10
24. Li, L., Gao, X., Deng, J., Tu, Y., Zha, Z.J., Huang, Q.: Long short-term relation transformer with global gating for video captioning. IEEE TIP **31**, 2726–2738 (2022) 4
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv (2023) 4
26. Lin, Z., Tan, C., Hu, J.F., Jin, Z., Ye, T., Zheng, W.S.: Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In: CVPR (2023) 2, 4, 10, 11, 14
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2024) 4
28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv (2019) 5, 10
29. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR (2022) 5, 10
30. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv (2023) 4
31. Medsker, L.R., Jain, L., et al.: Recurrent neural networks. Design and Applications **5**(64-67), 2 (2001) 4
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019) 10

33. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. In: NeurIPS. pp. 119336–119360 (2025) 4
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015) 4, 7
35. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2024) 4
36. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2024) 4
37. Su, R., Yu, Q., Xu, D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: ICCV (2021) 2, 3
38. Talal Wasim, S., Naseer, M., Khan, S., Yang, M.H., Shahbaz Khan, F.: Videogroundingdino: Towards open-vocabulary spatio-temporal video grounding. In: CVPR (2024) 4
39. Tan, C., Lin, Z., Hu, J.F., Li, X., Zheng, W.S.: Augmented 2d-tan: A two-stage approach for human-centric spatio-temporal video grounding. arXiv (2021) 3, 14
40. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. IEEE TCSVT **32**(12), 8238–8249 (2021) 2, 3, 10
41. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv (2023) 4
42. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE T-PAMI **40**(6), 1510–1517 (2017) 4
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017) 4, 6, 7, 8
44. Wang, Y., Xie, C., Liu, Y., Zheng, Z.: Videollamb: Long-context video understanding with recurrent memory bridges. arXiv (2024) 4
45. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: AAAI (2022) 3, 14
46. Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In: CVPR (2024) 2
47. Wu, C.Y., Krahenbuhl, P.: Towards long-form video understanding. In: CVPR (2021) 4
48. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR (2022) 4
49. Wu, Z., Xiong, C., Ma, C.Y., Socher, R., Davis, L.S.: Adaframe: Adaptive frame selection for fast video recognition. In: CVPR (2019) 4
50. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: CVPR (2022) 2, 11, 13, 14
51. Yu, Y., Wang, X., Hu, W., Luo, X., Li, C.: 2rd place solutions in the hc-stvg track of person in context challenge 2021. arXiv (2021) 3
52. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv (2024) 4
53. Zhang, Z., Zhao, Z., Lin, Z., Huai, B., Yuan, N.J.: Object-aware multi-branch relation networks for spatio-temporal video grounding. In: IJCAI (2020) 2
54. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: CVPR (2020) 1, 2, 3