LTCA: Long-range Temporal Context Attention for Referring Video Object Segmentation

Cilin Yan*, Jingyun Wang*, Guoliang Kang†

***Abstract*—Referring Video Segmentation (RVOS) aims to segment objects in videos given linguistic expressions. The key to solving RVOS is to extract long-range temporal context information from the interactions of expressions and videos to depict the dynamic attributes of each object. Previous works either adopt attention across all the frames or stack dense local attention to achieve a global view of temporal context. However, they fail to strike a good balance between locality and globality, and the computation complexity significantly increases with the increase of video length. In this paper, we propose an effective long-range temporal context attention (LTCA) mechanism to aggregate global context information into object features. Specifically, we aggregate the global context information from two aspects. Firstly, we stack sparse local attentions to balance the locality and globality. We design a dilated window attention across frames to aggregate local context information and perform such attention in a stack of layers to enable a global view. Further, we enable each query to attend to a small group of keys randomly selected from a global pool to enhance the globality. Secondly, we design a global query to interact with all the other queries to directly encode the global context information. Experiments show our method achieves new state-of-the-art on four referring video segmentation benchmarks. Notably, our method shows an improvement of 11.3% and 8.3% on the MeViS val^u and val datasets respectively. The code is available at https://github.com/cilinyan/LTCA.**

I. Introduction

Referring video object segmentation (RVOS) [1], [2] is a specialized form of segmentation task [3]–[15] that requires identifying and segmenting target objects in a video using detailed language expressions as guidance. RVOS is challenging as we need to build interactions between expressions and videos and align context information across modalities to localize and segment the desired object. Though challenging, RVOS is of vital importance in practice, as it is natural to depict our demand in language expressions and segment the desired objects in videos under the guidance of natural language. RVOS may benefit many real-world applications, *e.g.,* autonomous driving [16], [17], video surveillance [18], [19], video editing [20], [21], human-robot interaction [22], [23], *etc*. For instance, in human-robot interaction, through RVOS, users can instruct a robot to complete a specific task, which is important for seamless interaction and task completion in dynamic environments.

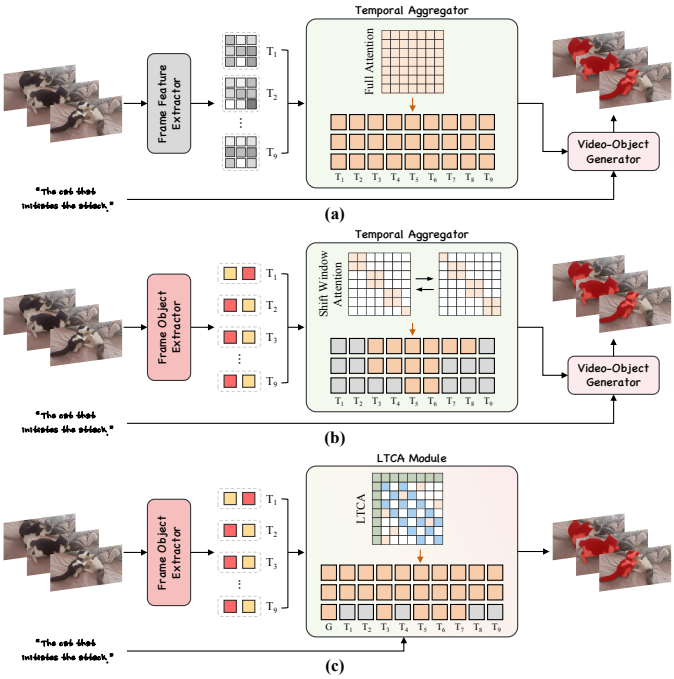


Fig. 1: Comparison of current query-based referring video segmentation (RVOS) pipelines. (a) Modeling frame feature sequence using full attention, (b) Modeling frame object sequence using shift window attention, (c) Modeling frame object sequence using LTCA.

The key to solving RVOS lies in extracting long-range temporal context information from the interactions between expressions and videos, and effectively capturing the dynamic attributes of each object. Among the long-range temporal context information, there exist both local and global information. For locality, we expect the model to emphasize local context. To realize this, each frame only need to attend to its neighbors within a certain temporal range to extract the context information. For globality, we expect the model to capture global temporal information across all frames in the video. Currently, there exist two typical ways to aggregate the long-range temporal context information in RVOS. One way (*e.g.,* [2] in Fig. 1(a)) is to adopt a full attention across all the frames, *i.e.,* each frame is capable of attending to all the other frames to aggregate useful information. However, for long sequences of frames, the magnitudes of full attention weights may become quite close, rendering it hard to emphasize local important context information. As a result, the full-attention way may capture a global view of temporal

*Equal contribution. †Corresponding author.
Cilin Yan, Jingyun Wang and Guoliang Kang are with School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: clyan@buaa.edu.cn, 19231136@buaa.edu.cn, kgl.prml@gmail.com)

context, but sacrifice important local information, exhibiting a bad locality. Another way is to apply dense local attention (*e.g.,* shift window attention [24] in Fig. 1(b)) to aggregate local context information and stack such attention layers to obtain a global context. However, as the length of video increases, in order to capture global temporal information, the number of attention layers should also increase, resulting in larger time and memory costs. Thus, in this way, although the locality is enhanced, the global context modeling ability is weakened. To summarize, previous works fail to strike a good balance between locality and globality, and the computation complexity may significantly increase with the increase of video length. Various attention mechanisms have been explored in action/motion recognition task [25]–[30]. However, in RVOS task, we need to model the long-range temporal context information for each spatial position and model the text-video interaction, the massive attention computation of which renders the attention mechanism proposed in previous action/motion recognition task not applicable to the RVOS task.

In this paper, we introduce an effective long-range temporal context attention (LTCA) mechanism for RVOS task. Specifically, the queries, which are initialized by the linguistic expression embeddings, interact with extracted frame features to encode the object and context information for each frame. LTCA (as illustrated in Fig. 1(c)) then aggregates the long-range temporal context information and manages to achieve a good balance between locality and globality with the following specific designs: Firstly, we design a dilated window attention mechanism across frames to aggregate local context information effectively and perform this attention over multiple layers to enable a comprehensive global view. Secondly, we design a random attention mechanism to enhance overall globality by allowing each query to attend to a small, randomly selected group of keys from a global pool. Thirdly, we introduce global queries to interact with all the other queries to encode the global context information and further improve the globality. Thanks to LTCA, we may directly compare aggregated global query features and the frame feature maps to generate the object masks across frames, without the need to rely on a video-object generator to correlate expressions and aggregated features (as shown in Fig. 1(a-b)).

Extensive experiments are conducted on representative benchmarks including MeViS [24], Ref-YouTube-VOS [31], Ref-DAVIS17 [32] and A2D-Sentences [33]. Our method achieves state-of-the-art performance on four benchmarks. Note that among these benchmarks, MeViS contains more motion expressions, which proposes a higher demand for utilizing temporal information. Our superior performance on MeViS (outperforming previous state-of-the-art by 11.3%) verifies the effectiveness of LTCA to aggregate long-range temporal information.

In summary, our contributions are as follows:

- We propose a new attention mechanism LTCA to capture the long-range temporal context for RVOS. LTCA stacks sparse local attention (including dilated window attention and random attention) layers to balance locality and globality, and directly aggregates global information via specifically designed global queries in global attention.
- Based on LTCA, we propose a simplified framework, where queries are firstly encoded through interactions between expressions and frame features, then object queries aggregate temporal context information through LTCA, and finally the object masks across frames are generated by comparing global queries with the frame features.
- Extensive experiments on four representative benchmarks (*i.e.,* MeViS, Ref-YouTube-VOS, Ref-DAVIS17, A2D-Setences) show that our method outperforms previous works remarkably, achieving new state-of-the-art.

II. Related Work

In this section, we first introduce the related works of video instance segmentation. Afterward, we give a brief review of referring image segmentation. Finally, we introduce referring video object segmentation.

A. Video Instance Segmentation

Video instance segmentation (VIS) [34]–[43] is a joint vision task, which requires to identify, segment and track the interested instances in a video simultaneously. Existing VIS methods can be mainly divided into two categories: online and offline approaches. Online methods [34]–[38] process a video frame-by-frame, segmenting objects in a frame level and associating instances across time. MaskTrack R-CNN [37] incorporates a tracking head based on Mask R-CNN [44] and associates instances in different frames by heuristic cues. Following the recent success of Transformer [45], several Transformer-based methodologies have been introduced. For instance, MinVIS [34] recognizes the discriminative nature of query embeddings and employs a straightforward strategy by exclusively aligning these embeddings between the current and adjacent frames. Meanwhile, CTVIS [38] learns more robust and discriminative instance embeddings by aligning the training and inference pipelines. Offline methods [46]–[48] take the entire video as input, performing segmentation on all frames and associating instances at one time. SeqFormer [47] processes spatio-temporal features and directly predicts instance mask sequences, while VITA [48] proposes to associate frame-level object tokens without using dense spatio-temporal backbone features and reduce the memory consumption on long videos.

B. Referring Image Segmentation

Referring Image Segmentation (RIS) [49]–[51] aims to segment the target object in the given static image referred by language. Early studies [52], [53] have adhered to a concatenate-then-convolve approach, in which the fusion of language and image features is accomplished through concatenation. Subsequent works [54], [55] have enhanced segmentation performance by employing RNN or dynamic convolution techniques, or by investigating optimal position for executing language-image fusion. Building on the success of attention architecture [56], current studies [57]–[59] primarily employ

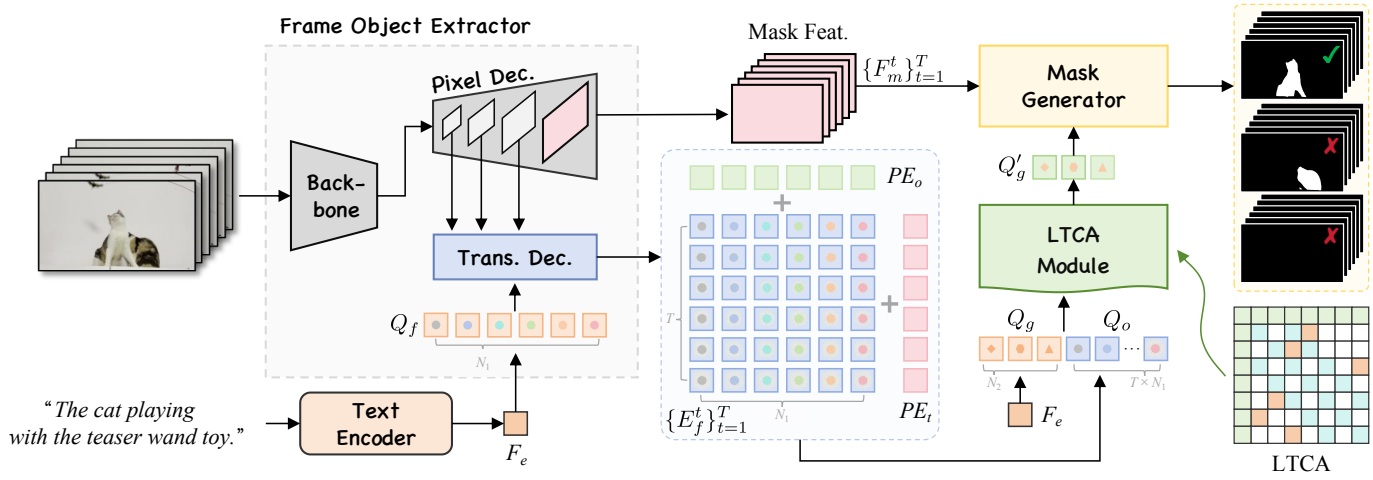


Fig. 2: **The overview architecture of our method.** First, all input frames are fed to a transformer-based extractor to generate object-centric embeddings $\{E_f^t\}_{t=1}^T$. Then we flatten $\{E_f^t\}_{t=1}^T$ as frame object queries Q_o and concatenate it with a set of learnable global queries Q_g, which is initialized with text embedding of given linguistic expressions. Then the concatenated queries are fed to an LTCA module to conduct efficient information interaction among frame object queries Q_o and linguistic-aware global queries Q_g. The output global queries $\widetilde{Q}_g$ are adopted to generate segmentation mask sequences of target objects.

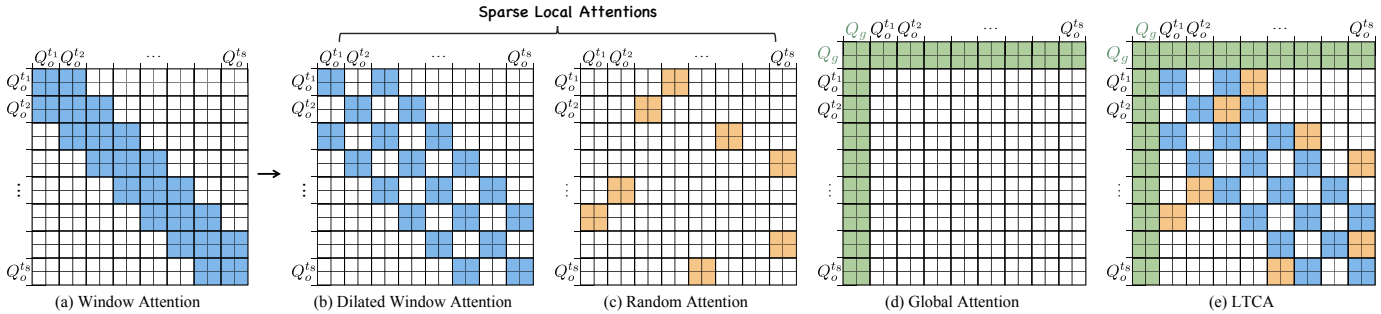
cross-attention modules for language-image fusion. For instance, LAVT [59] integrates the fusion of visual and linguistic features using unidirectional cross-attention at the intermediate layers of encoders. Additionally, VPD [60] attempts to harness semantic information in diffusion models for RIS, while LISA [61] and PixelLM [62] draws on the language generation capabilities of multimodel large language models [63], [64] to produce a segmentation mask in response to a complex and implicit query text. SESAME [65] introduces supplementary data to resolve the problem where LISA may be misled by false premises when queries imply the existence of objects that are not actually present in the image.

C. Referring Video Object Segmentation

Referring Video Object Segmentation (RVOS) is a challenging multi-modal video task, which aims to segment the target object indicated by a given language expression across the entire video. Most existing benchmarks of RVOS, including Ref-YouTube-VOS [31], Ref-DAVIS 2017 [32] and A2D-Sentences [33], primarily focus on salient objects, utilizing language descriptions of static attributes. MeViS [24] is a large-scale video dataset that contains videos with longer duration and expressions referring to an object with highly dynamic features, such as movements and state changes. A robust capability for long-range temporal feature modeling is essential for tackling MeViS dataset, but most existing methods [1], [2], [24], [31], [66] lack this capability. MTTR [1] adopts a query-based end-to-end framework to decode objects from multi-modal features frame by frame independently, thus failing to model the inter-frame temporal information. ReferFormer [2] employs full attention on frame-level tokens to model temporal information. However, the overly long frame features and quadratic complexity of full attention results in vast time and memory costs. Moreover, for input sequences of long-duration videos, the distribution of attention weight becomes smooth, making the model only focus on global context information but fail to effectively model the local temporal information. R^{2}VOS [67] began by augmenting the dataset with mismatched video-text pairs. It then utilized a contrastive learning approach to identify and eliminate these mismatches, thereby refining the model's grasp of language features. TempCD [68] aligned the natural language expression with the objects' motions and their dynamic associations at the global video level and segmented objects at the frame level. DMFormer [69] enhances explicit and discriminative feature embedding and alignment by introducing lightweight subject-aware and context-aware branches. HTR [70] addresses the challenge of temporal instance consistency by introducing a hybrid memory framework with inter-frame collaboration and a Mask Consistency Score metric. Sun et al. [71] propose a unified framework for multi-modality VOS, employing self-attention to process inputs as tokens and using reinforcement learning to optimize filter networks for highlighting significant information. However, [68] and [72] utilize the full attention mechanism, with the computational complexity of $O(n^2)$, rendering them impractical for long video scenarios. Recent works [24], [73] exploits a shift-window attention mechanism in the Transformer architecture, motivated by VITA [48]. Although the shift-window attention reduces the inference cost, it only models the local context information, losing the ability to model global context information for long videos. In this study, we introduce an efficient long-range temporal context attention mechanism that strikes a balance between locality and globality, while simultaneously simplifying the framework of RVOS.

III. Method

In this work, we propose a new method for Referring Video Segmentation task. RVOS aims to segment target objects in all frames of the video according to a given linguistic reference. The visual input is a sequence of video frames and the textual input is a sentence describing specific objects.



(a) Window Attention (b) Dilated Window Attention (c) Random Attention (d) Global Attention (e) LTCA

Fig. 3: Visualization of different attention patterns. For ease of visualization, we set $N_1 = 2$, $N_2 = 2$, $w = 3$, $d = 2$ and $r = 1$.

The general framework of our method is illustrated in Fig. 2. Given a language expression containing L words as text input, we encode it in word level by an off-the-shelf text encoder of RoBERTa [74]. We then obtain a sentence-level text feature F_e by averaging the obtained word embeddings. Object queries Q_f are initialized by repeating F_e for N_1 (number of potential objects per frame) times and global queries Q_g are initialized by repeating F_e for N_2 (total number of potential objects in the whole video) times. Passing the video frames $V = \{I^t\}_{t=1}^T$ and object queries $\{Q_f^t\}_{t=1}^T$ to the Frame Object Extractor $\mathcal{E}_f$ (Sec. III-B), mask features $\{F_m^t\}_{t=1}^T$ and embeddings of N_1 potential objects $\{E_f^t\}_{t=1}^T$ in each frame are generated. We then flatten the object embeddings $\{E_f^t\}_{t=1}^T$ as frame object queries Q_o and concatenate them with the linguistic-aware global queries Q_g to form the input of the LTCA (Long-range Temporal Context Attention) Module $\mathcal{E}_v$. Inside the LTCA Module, we propose a new LTCA mechanism (Sec. III-C), consisting of Global Attention and Sparse Local Attention, to effectively aggregate global context information into object features and keep a good balance between globality and locality. Therefore, each query in $\widetilde{Q}_g$ processed by LTCA module represents an object instance proposal within the video regarding the linguistic expression. Sending these global queries $\widetilde{Q}_g$ to Mask Generator (Sec. III-D), we finally generate a sequence of target masks $\{M_i\}_{i=1}^{N_2}$.

A. Baseline: Segment via Shift Window Attention

LMPM [24] performs RVOS task by aggregating temporal information across the entire video via shift-window attention mechanism. LMPM obtains mask features $\{F_m^t\}_{t=1}^T$ and object embeddings $\{E_f^t\}_{t=1}^T$ from Frame Object Extractor. Flattening frame-level object queries $\{E_f^t\}_{t=1}^T$ as frame object queries Q_o, LMPM passes it to the temporal aggregator with shifted window attention mechanism to obtain contextual temporal information $\hat{Q}_o$ across T frames. Subsequently, the Transformer decoder treats the linguistic-aware global queries Q_g as queries and uses the obtained object embeddings $\hat{Q}_o$ as keys and values, to predict the masks and confidence scores of the target objects.

Although LMPM enhances the locality, the global context modeling is weakened. For example, given the length of window w_s and the number of layer k in shift-window attention mechanism, two frames can contact with each other only if the distance between them is less than $\Theta(kw_s/2)$. In other words, as the length of the target video increases, in order to capture global temporal information of video, the number of attention layers should also increase, thus leading to larger time and memory costs. Besides, the last Transformer decoder is unnecessary since the frame-level object queries Q_o can be simply encoded into linguistic-aware global queries Q_g in the object encoder. Based on the observations above, we propose a new framework with a newly designed LTCA module in the following sections.

B. Frame Object Extractor

Mask2Former [75] is adopted here as a frame object Extractor $\mathcal{E}_f$. A sequence of T video frames $V = \{I^t\}_{t=1}^T$ are sent into the extractor as visual input and frame-level low-resolution image features are extracted by the visual backbone. A pixel decoder gradually upsamples these low-resolution features to form multi-scale image features. For one branch, the final high-resolution per-pixel image features are treated as mask features $\{F_M^t\}_{t=1}^T$ for further mask generation. For the other branch, image features of different scales are sent into a Transformer decoder as keys and values for different layers successively and are decoded through cross-attention with the object queries $\{Q_f^t\}_{t=1}^T$, yielding object embeddings $\{E_f^t\}_{t=1}^T$. The process can be written as follows,

$$\{F_M^t\}_{t=1}^T, \{E_f^t\}_{t=1}^T = \mathcal{E}_f(\{I^t\}_{t=1}^T). \tag{1}$$

As a result, each potential object in each frame is represented by an output object query. Then, the output frame-level object queries are used to obtain global temporal information in the Long-range Temporal Context Attention mechanism and generate linguistic-guided segmentation kernels in Mask Generator.

C. Long-range Temporal Context Attention (LTCA) Module

1) ***LTCA Module****:* Long-range Temporal Context Attention (LTCA) Module $\mathcal{E}_v$ simultaneously aggregates long-range temporal context information and models video-level object representations with linear computational complexity. LTCA Module consists of a stack of our proposed Transformer encoders as modified versions of the standard ones. The key components of our Transformer encoder include: a) global queries, which extract video-level object representations, b)

masked attention operator. [1], which models videos with linear computational complexity.

We flatten the object embeddings $\{E_t^t\}_{t=1}^T \in \mathbb{R}^{T\times N_1\times D}$ as frame object queries $Q'_o \in \mathbb{R}^{TN_1\times D}$, where D is the feature dimension of object embeddings. We then combine learnable object position embeddings $PE_o \in \mathbb{R}^{N_1\times D}$ with non-learnable sinusoidal frame positional embeddings $PE_t \in \mathbb{R}^{T\times D}$, and then add them to frame object queries Q'_o to obtain $Q_o \in \mathbb{R}^{TN_1\times D}$ as input for the encoder, which can be represented as:

$$Q_o^{(i-1)\times N_1+j} = Q_o^{'(i-1)\times N_1+j} + PE_o^j + PE_t^i, \tag{2}$$

where $i \in \{1, 2, ..., T\}$ and $j \in \{1, 2, ..., N_1\}$. We then concatenate the frame object queries Q_o with linguistically-aware global queries $Q_g \in \mathbb{R}^{N_2\times D}$ to form the input Q_v for the LTCA Module $\mathcal{E}_v$, illustrated as follows:

$$Q_v = [Q_g, Q_o] \in \mathbb{R}^{N_2+TN_1}. \tag{3}$$

The attention operation of LTCA module can be expressed as

$$Q_v^l = \mathrm{softmax}(M^{l-1} + Q^l K^{l\,T})V^l + Q_v^{l-1}. \tag{4}$$

Here, l is the layer index of LTCA module, Q_v^l refers to $N_2 + TN_1$ D-dimensional query features at the l^{th} layer and $Q^l, K^l, V^l \in \mathbb{R}^{(N_2+TN_1)\times D}$ are the query features under linear transformation $f_Q(\cdot)$, $f_K(\cdot)$ and $f_V(\cdot)$ respectively. Q_v^0 is the input query features Q_v. Moreover, M^{l-1} denote the attention mask for the $(l-1)^{\mathrm{th}}$ layer. As illustrated in Fig. 3(e), the positions that are non-white in M^{l-1} are filled with zeros, while all other colored positions are filled with $-\infty$. The acquisition of M^{l-1} will be described in Sec. III-C2. Through LTCA module, we can derive video-level query representations, denoted as $\widetilde{Q}_g$, which can be written as,

$$[\widetilde{Q}_g, \widetilde{Q}_o] = Q_v^{-1} = \mathcal{E}_v(Q_v) = \mathcal{E}_v([Q_g, Q_o]). \tag{5}$$

*2) **LTCA mechanism**:* Our LTCA module uses Long-range Temporal Context Attention (LTCA) to capture long-range temporal context for RVOS. LTCA is designed to aggregate temporal context information in two ways: one is to balance the locality and globality via stacking sparse local attention layers, and the other one is to directly aggregate the global information.

Sparse Local Attention Firstly, we introduce window attention across frames to aggregate local context information. As shown in Fig. 3(a), the embedding of the current frame can attend to the embeddings of its w neighboring frames on both sides, which is beneficial for the construction of local contextual temporal information. As shown in Eq. 4, the forms of different attention operations differ only in the definition of attention mask M. With $\phi(i)$ representing the index of the video frame where query i is extracted, the window attention mask M_{ij} with respect to query i is defined as

$$M_{ij} = \begin{cases} 0, & \text{if } |\phi(i) - \phi(j)| \le \dfrac{w}{2} \\ -\infty, & \text{otherwise} \end{cases}. \tag{6}$$

[1]For explanatory purposes, we use an attention mask to illustrate the Long-range Temporal Context Attention mechanism. We achieve linear complexity computation by performing roll operations on keys.

Note that among all mask definitions of local attention, i and j are not in the global query set. From Eq. 4, $M_{ij} = 0$ means query i can attend to query j normally. In contrast, when $M_{ij} = -\infty$, query i cannot attend to query j as the corresponding attention weight approximates zero. However, the receptive field of window attention, which is a form of dense local attention, is limited and fails to strike a good balance between locality and globality. Therefore, we further introduce dilated window attention to address this issue inspired by dilated convolutions [76]. As shown in Fig. 3(b), dilated window attention is an extension of window attention, where the current frame attends to its neighboring frames at an interval of d frames within a certain window range. Window attention can also be viewed as dilated window attention with $d = 1$. The computational complexity of dilated window attention is $O(wN_1T)$, which is independent of d. The introduction of dilated window attention, a form of sparse local attention, effectively expands the receptive field of each frame, thereby enabling better aggregation of a relatively larger amount of global information. Dilated window attention mask is defined as

$$M_{ij} = \begin{cases} 0, & \text{if } |\phi(i) - \phi(j)|\%d = 0, |\phi(i) - \phi(j)| \le \dfrac{dw}{2} \\ -\infty, & \text{otherwise} \end{cases}. \tag{7}$$

To further enhance the global information aggregation capability of LTCA, we introduce random attention, inspired by ShuffleNet [77]. As shown in Fig. 3(c), random attention enables each frame to attend to r frames randomly selected from the entire video, which means that the receptive field of each frame is no longer limited, thereby enhancing globality. Although random attention introduces randomness into inference, our experiments show that this randomness does not have a significant impact on inference (the standard deviation of metrics is less than 0.2%). With $\psi_{\phi(i)}(T, r)$ denoting the randomly-selected r frames from the set $\{1, 2, \ldots, T\}$ for $\phi(i)$, we define the random attention mask as

$$M_{ij} = \begin{cases} 0, & \text{if } \phi(j) \in \psi_{\phi(i)}(T, r) \\ -\infty, & \text{otherwise} \end{cases}. \tag{8}$$

Global Attention Although stacking sparse local attention can enable a global view of temporal context, sparse local attention may still not be able to fully capture the useful global context information. Therefore, we further introduce global attention, which uses additionally introduced global queries to directly aggregate global context information. We initialize the global queries with language features to regularize their training. As shown in Fig. 3(b), global attention introduces N_2 additional queries Q_g, where all object queries Q_o attend to global queries Q_g, and simultaneously, global queries Q_g attend to the remaining object queries Q_o. This operation ensures that any two queries can exchange information as long as two Transformer layers are stacked, solving the problem that the previous linear computational complexity attention could not fully model long videos. Global attention is a sparse attention mechanism, and its computational complexity is $O(2N_2N_1T)$,

TABLE I: Performance of our method on MeViS Datasets. † denotes the LMPM model retrained using our training parameters.

Methods	Pub.&Year	Backbone	val^u			val		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [31]	ECCV 2020	Swin-T	-	-	-	27.8	25.7	29.9
LBDT [78]	CVPR 2022		-	-	-	29.3	27.8	30.8
MTTR [1]	CVPR 2022		-	-	-	30.0	28.8	31.2
ReferFormer [2]	CVPR 2022		-	-	-	31.0	29.8	32.2
VLT+TC [79]	TPAMI 2023		-	-	-	35.5	33.6	37.3
LMPM [24]	ICCV 2023		40.2	36.5	43.9	37.2	34.2	40.2
LMPM†	ICCV 2023		48.3	43.9	52.6	41.8	38.7	44.9
Ours			**51.5±0.1**	**46.6±0.2**	**56.4±0.1**	**45.3±0.1**	**41.7±0.1**	**48.9±0.1**
Ours		Swin-S	52.9±0.2	47.8±0.2	58.1±0.3	45.9±0.1	42.0±0.1	49.7±0.0
Ours		Swin-B	55.4±0.1	50.8±0.1	60.0±0.1	46.6±0.1	43.2±0.1	50.0±0.1
Ours		Swin-L	56.0±0.1	51.0±0.1	61.1±0.2	47.5±0.1	43.5±0.1	51.4±0.1

which is still linear. In global attention, the global queries aggregate Q_g the contextual temporal information of each instance throughout the video, and the object queries of each frame attend to other objects through global queries as a bridge. Specifically, global attention mask can be expressed as

$$M_{ij} = \begin{cases} 0, & \text{if } i \in \mathcal{G} \text{ or } j \in \mathcal{G} \\ -\infty, & \text{otherwise} \end{cases}, \tag{9}$$

where $\mathcal{G}$ represents the global query set.

Furthermore, in LTCA, global queries Q_g can be utilized to directly predict the object mask and confidence score as they have effectively gathered long-range temporal context information of each instance in the video with a good balance between locality and globality. Therefore, the introduction of global attention negates the need for an additional video-object generator (*i.e.,* Transformer Decoder in LMPM [24]) to construct global instance embeddings, further simplifying the framework for RVOS task.

Discussion Sparse local attention (including dilated window attention and random attention) only attends to frames within a limited temporal range to maintain good locality. By stacking them layer by layer, a global view can be achieved. As sparse local attention mainly aims to enhance the locality, we design global queries to directly aggregate global context information. Thus, by combining sparse local attention and global attention, LTCA achieves a good balance between locality and globality.

D. Mask Generator

After the LTCA module, the global queries $\widetilde{Q}_g$ aggregate the temporal information of the entire video and represent several instances within the video. We feed the global queries and mask features into the Mask Generator to obtain the final predicted masks and confidence scores. Specifically, the Mask Generator comprises of the segmentation head $\mathcal{H}_s$ and the classification head $\mathcal{H}_c$. The segmentation head $\mathcal{H}_s$ containing three MLP layers is applied to generate the set of segmentation masks $\{M_i\}_{i=1}^{N_2} \in \mathbb{R}^{T\times N_2\times \frac{H}{S}\times \frac{W}{S}}$, and each mask M_i^t of the t^{th} frame is obtained by

$$M_i^t = F_m^t \otimes \mathcal{H}_s(\widetilde{Q}_{g_i}), \tag{10}$$

where the $\mathcal{H}_s(\widetilde{Q}_{g_i})$ indicates the dot product weight generated by global query of the last Transformer encoder layer, and the $\otimes$ means dot product operation. With the segmentation head, each global query generates segmentation masks of T frames. Meanwhile, the global queries $\widetilde{Q}_{g_i}$ are concatenated with the language feature F_e and inputted into the classification head $\mathcal{H}_c$, which consists of a single linear layer, to obtain the prediction confidence scores $\{S_i\}_{i=1}^{N_2} \in \mathbb{R}^{N2}$ for prediction masks, which can be represented as follows:

$$S_i = \mathcal{H}_c(\widetilde{Q}_{g_i}). \tag{11}$$

Following LMPM [24], we use $\mathcal{L}_f$ from [75] to calculate loss from the per-frame outputs to frame-wise ground-truth, use $\mathcal{L}_{sim}$ and $\mathcal{L}_v$ from [48] to calculate the video-level loss. Finally, we integrate all losses together as follows:

$$\mathcal{L}_{total} = \lambda_v\mathcal{L}_v + \lambda_f\mathcal{L}_f + \lambda_{sim}\mathcal{L}_{sim}. \tag{12}$$

During the inference stage, for RVOS tasks that predict a single mask output, such as the Ref-YouTube-VOS dataset [31], we select the mask with the highest prediction score as the final prediction result. For RVOS tasks that predict multiple mask outputs, such as the MeViS dataset [24], we select the masks with prediction confidence scores greater than the confidence threshold σ as the final prediction results.

IV. Experiments

A. Dataset and Evaluation Metrics

The proposed methods are evaluated on four video segmentation datasets: MeViS [24], Ref-YouTube-VOS [31], Ref-DAVIS17 [32] and A2D-Sentences [33]. Following [24], we use region similarity metric $\mathcal{J}$, F-measure $\mathcal{F}$ and the average of these two metrics $\mathcal{J}\&\mathcal{F}$ as the evaluation metrics on MeViS Ref-YouTube-VOS and Ref-DAVIS17. Following [1], we use Overall IoU, Mean IoU and MAP over 0.50:0.05:0.95 as the evaluation metrics on A2D-Sentences. Following [2], [73], on Ref-DAVIS17, we directly report the results using the model trained on Ref-YouTube-VOS without finetuning. For more detailed descriptions of the datasets and implementation details, please refer to the Appendix.

TABLE II: Performance of our method on Ref-YouTube-VOS and Ref-DAVIS17 Datasets. Following conventional practice, "With Image Pretraining" means the models are first pretrained on RefCOCO [80], RefCOCO+ [80], and RefCOCOg [81] datasets. "Joint Training" means the models are trained with the combination of image datasets and video datasets.

Methods	Pub.&Year	Backbone	Ref-YouTube-VOS			Ref-DAVIS17		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
With Image Pretraining								
URVOS [31]	ECCV2020	ResNet-50	47.2	45.3	49.2	51.5	47.3	56.0
MLRL [82]	CVPR 2022		49.7	48.4	51.0	52.7	50.1	55.4
ReferFormer [2]	CVPR 2022		55.6	54.8	56.5	58.5	55.8	61.3
OnlineRefer [83]	ICCV 2023		57.3	55.6	58.9	59.3	55.7	62.9
TempCD [68]	ICCV 2023		59.0	57.5	60.5	60.0	57.3	62.7
Ours			**59.7±0.2**	**58.1±0.1**	**61.3±0.2**	**60.6±0.1**	**57.4±0.1**	**63.7±0.1**
ReferFormer [2]	CVPR 2022	Video-Swin-T	59.4	58.0	60.9	59.7	56.6	62.8
TempCD [68]	ICCV 2023		62.3	60.5	64.0	62.2	59.3	65.0
SOC [72]	NeurIPS 2023		62.4	61.1	63.7	63.5	**60.2**	66.7
Ours			**62.9±0.1**	**61.5±0.1**	**64.3±0.1**	**63.6±0.1**	60.1±0.1	**67.1±0.1**
Joint Training								
CMSA [84]	CVPR 2019	ResNet-50	34.9	33.3	36.5	34.7	32.2	37.2
LBDT-4 [78]	CVPR 2022		48.2	50.6	49.4	-	-	-
YOFO [85]	AAAI 2022		48.6	47.5	49.7	53.3	48.8	57.9
ReferFormer [2]	CVPR 2022		58.7	57.4	60.1	61.1	58.0	64.1
UniRef [86]	CVPR 2023		60.6	59.0	62.3	-	-	-
MUTR [73]	AAAI 2024		61.9	60.4	63.4	65.3	62.4	68.2
Ours			**62.5±0.2**	**60.6±0.1**	**64.4±0.2**	**66.0±0.1**	**62.8±0.0**	**69.3±0.1**
ReferFormer [2]	CVPR 2022	Swin-L	64.2	62.3	66.2	63.9	60.8	67.0
UniRef [86]	CVPR 2023		67.4	65.5	69.2	-	-	-
MUTR [73]	AAAI 2024		68.4	66.4	70.4	68.0	64.8	71.3
Ours			**69.1±0.1**	**67.1±0.1**	**71.2±0.0**	**68.6±0.1**	**65.2±0.1**	**71.9±0.1**

TABLE III: Performance of our method on A2D-Sentences.

Method	Pub.&Year	Backbone	IoU		mAP
			Overall	Mean	
ACAN [87]	ICCV 2019	I3D	60.1	49.0	27.4
CMPC-V [88]	TPAMI 2021	I3D	65.3	57.3	40.4
MTTR [1]	CVPR 2022	Video-Swin-T	72.0	64.0	46.1
ReferFormer [2]	CVPR 2022	Video-Swin-T	72.3	64.1	48.6
SOC [72]	NeurIPS 2023	Video-Swin-T	74.7	66.9	50.4
Ours		Video-Swin-T	**75.5**	**67.7**	**51.4**

B. Comparison with previous state-of-the-arts

We compare our method to previous state-of-the-art RVOS methods in Table I and Table II. The numbers reported in the tables are directly cited from corresponding papers, except for LMPM† which is retrained using our training setups.

MeViS We compare our method to previous works on MeViS benchmark in Table I. Our approach outperforms all the previous works remarkably, *e.g.,* outperforming LMPM by around 11.3% and 8.1% $\mathcal{J}\&\mathcal{F}$ on MeViS val^u and val datasets. As MeViS contains more motion expressions than the other two benchmarks, the superior performance of our method verifies the effectiveness of our LTCA in aggregating temporal information.

Ref-YouTube-VOS & Ref-DAVIS17 We compare our method to previous models on Ref-YouTubeVOS and Ref-DAVIS17 in Table II. Our method achieves new state-of-the-art performance among different training settings: with image pretraining, and joint training. For example, our method outperforms MUTR by about 0.7% $\mathcal{J}\&\mathcal{F}$ on Ref-YouTube-VOS dataset with Swin-L backbone. The less significant improvement on Ref-YouTube-VOS and Ref-DAVIS17 may be attributed to the fact that these datasets primarily focus on static objects with fewer language expressions depicting motion. In contrast, our method concentrates on long-range temporal context modeling.

A2D-Sentences As shown in Table III, our method outperforms the current state-of-the-art methods, SoC [72] and ReferFormer [2]. When trained from scratch, our model achieves 51.4% mAP and 67.7% mIoU, with an improvement of 1.0% mAP and 0.8% mIoU compared to SOC.

C. Ablation Studies

Effect of Each Component. In Table IV, we conduct experiments to verify the effectiveness of each key component in our LTCA. When combining the global attention with the shift window attention (adopted by LMPM), the performance improves 2.2% $\mathcal{J}\&\mathcal{F}$ on the val^u set compared with baseline model LMPM†, which verifies the effectiveness of our proposed global attention. By comparing the first two rows of the table, we observe that introducing the global attention may avoid using the Video-Object Generator. From the last three rows of the table, replacing the dense local attention (*i.e.,* shift window attention) with each kind of sparse local attention (*i.e.,* dilated window attention and random attention) yields an obvious improvement, verifying the effectiveness of our proposed sparse local attention. Both the global attention and the sparse local attention contribute to the final video object segmentation performance.

Effect of Video Length. a) In Table V(a), we split the MeViS val set into short and long video groups based on the number

TABLE IV: Effect of each attention component of LTCA on MeViS.

Global Attention	Sparse Local Attention		Video-Object Generator	val^u			val		
	Dilated	Random		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✓			✓	50.5	45.9	55.1	42.8	39.5	46.0
✓			✘	50.6	45.5	55.6	44.0	40.3	47.8
✓	✓		✘	51.0	45.9	56.1	44.3	40.4	48.1
✓		✓	✘	51.2±0.1	46.3±0.1	56.0±0.0	44.8±0.0	41.2±0.0	48.5±0.0
✓	✓	✓	✘	51.5±0.1	46.6±0.2	56.4±0.1	45.3±0.1	41.7±0.1	48.9±0.1

TABLE V: Comparison of metrics on long videos and short videos on the MeViS dataset. $\Delta\mathcal{J}\&\mathcal{F}$ represents the percentage change of metrics on long videos relative to short videos.

(a) We divided the MeViS val set into short videos and long videos, trying to equalize the number of long videos and short videos.

Method	Video			val			$\Delta\mathcal{J}\&\mathcal{F}$
	Type	Length	Number	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	(relative)
LMPM	Short	≤ 46	1121	38.0	34.9	41.0	-4.1%
	Long	> 46	1115	36.4	33.5	39.3	
LMPM†	Short	≤ 46	1121	41.9	38.9	44.9	-0.5%
	Long	> 46	1115	41.7	38.5	44.9	
Ours	Short	≤ 46	1121	44.9	41.5	48.3	+2.2%
	Long	> 46	1115	45.9	42.0	49.7	

(b) We selected videos where information exchange can occur between any two frames under the LMPM shift window attention mechanism as short videos, and the rest as long videos, with the video threshold set at 18.

Method	Video			val			$\Delta\mathcal{J}\&\mathcal{F}$
	Type	Length	Number	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	(relative)
LMPM	Short	≤ 18	21	56.9	55.5	58.3	-34.9%
	Long	> 18	2215	37.0	34.0	40.0	
LMPM†	Short	≤ 18	21	63.9	62.6	65.2	-34.9%
	Long	> 18	2215	41.6	38.5	44.7	
Ours	Short	≤ 18	21	57.8	56.3	59.3	-21.7%
	Long	> 18	2215	45.3	41.6	48.9	

of video frames. The results show that our method, performs better on long videos than on short ones, with an absolute advantage of 1% and a relative advantage of 2.2% on $\mathcal{J}\&\mathcal{F}$. In contrast, LMPM performs relatively worse on long videos than short videos. The comparisons show that compared to LMPM which stacks dense local attentions (*i.e.,* shift window attention), our LTCA is more effective to aggregate long-range temporal information.

b) As we know, LMPM ensures a global view by staking multiple shift window attention layers. However, with a certain number of layers, only partial frames can exchange their information, *i.e.,* the temporal receptive field L of each frame in the last layer relative to the first layer may be smaller than the video length. We use L to separate short and long videos, *i.e.,* the videos whose length is larger than L are treated as long videos, and those whose length is smaller than L are viewed as short videos. The results in Table V(b) show that under this division method, the performance loss of our method on long videos relative to short videos is lower than that of LMPM, further validating the superior ability of our method to model long videos. The reason we use sample quantity as the basis for division in Table V(a) is that text descriptions with video frames less than or equal to 18 in the MeViS val set only account for 1% of the val set, which may lead to less credible results. However, the comparisons under both division methods show that our method is superior to LMPM in modeling long videos.

Effect of Sparse Local Attention hyper-parameters. In Table VI, we conduct experiments on various hyper-parameters in sparse local attention of LTCA, including the number of attending neighbors w, dilated interval d and the number of randomly selected frames r respectively. From the table, we

TABLE VI: Effect of different sparse local attention setups on MeViS. The w, r and d are determined sequentially.

w	r	d	val		
			$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	0	1	43.5	39.6	47.4
3	0	1	**44.0**	**40.3**	**47.8**
5	0	1	43.9	40.4	47.3
3	0	1	44.0	40.3	47.8
3	1	1	44.3±0.1	40.5±0.1	48.1±0.0
3	2	1	**44.8±0.0**	**41.2±0.0**	**48.5±0.0**
3	3	1	44.6±0.2	40.9±0.2	48.3±0.1
3	2	1	44.8±0.0	41.2±0.0	48.5±0.0
3	2	2	**45.3±0.1**	**41.7±0.1**	**48.9±0.1**
3	2	3	44.1±0.1	40.5±0.1	47.7±0.1

observe that the optimal choices for w, r and d should neither be too large nor too small. We speculate it is because the optimal choices reach a good balance between locality and globality.

Comparison with Full Attention We firstly replace the sparse local attention in LTCA with full attention and train the model on MeViS. According to results in Table VII, training with dense global attention (Line 1) is inferior than LTCA (Line 3). Furthermore, we train a model with a combination of full attention and LTCA, which means a layer-wise alternation between the two attention mechanisms within the transformer encoder (LTCA module). We observe that such a combination (Line 2) also results in inferior results compared with LTCA.

Efficiency of Our Method. In Table VIII, we compare FLOPs, parameters, inference speed and GPU memory of our method with ReferFormer, MUTR and LMPM. As we avoid using a

TABLE VII: Comparison of LTCA and full attention on the MeViS dataset.

Method	val^u			val		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Full Attn.	50.8	45.8	55.7	42.6	39.8	45.4
Full Attn. + LTCA	51.1	46.2	56.0	44.3	40.8	47.8
LTCA	51.5	46.6	56.4	45.3	41.7	48.9

TABLE VIII: **Efficiency comparison with state-of-the-art methods.** The shape of the input videos is [48, 3, 448, 796], and the length of input text tokens is 40.

Methods	GFLOPs	Parameters	Infer. Time	GPU Memory
ReferFormer	7172	**176.4M**	1.39s	25540M
MUTR	7220	186.4M	1.42s	25641M
LMPM	4430	191.0M	0.98s	8165M
Ours	**4429**	186.3M	**0.97s**	**8121M**

Fig. 4: **Visualization Results on MeViS.** We compare the visualization results between LMPM and our method on MeViS. Previous work easily segments irrelevant objects, as this method fails to strike a good balance between locality and globality.

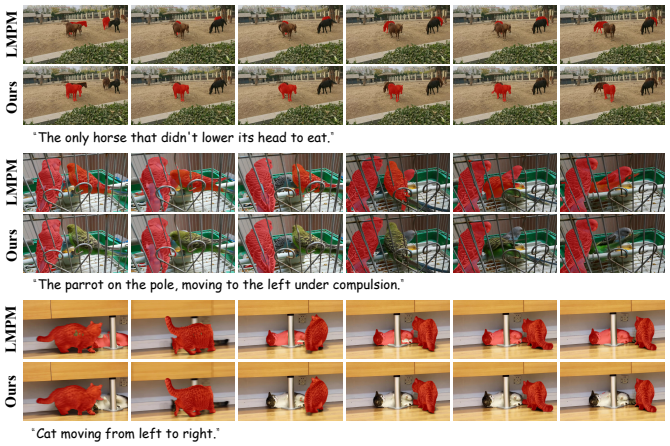


Fig. 5: **Visualization Results on Ref-YouTube-VOS.** We compare the visualization results between MUTR and our method on Ref-YouTube-VOS benchmarks.

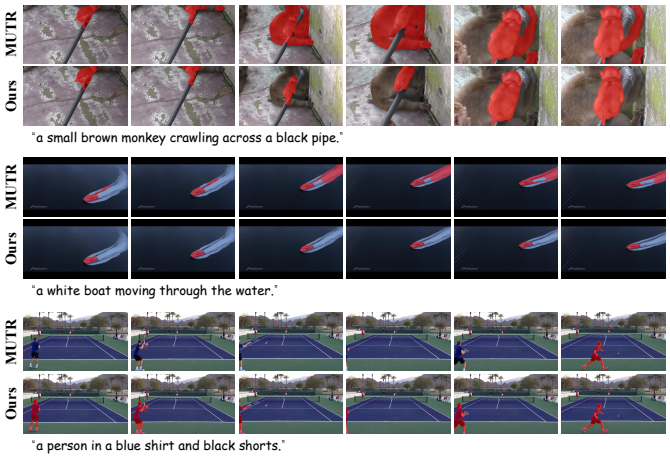


TABLE IX: Performance comparison with varying video lengths on the MeViS dataset.

Video		val		
Length	Number	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
$15 \leq L < 35$	35	43.7±0.2	40.9±0.2	46.5±0.2
$35 \leq L < 48$	35	45.7±0.1	41.8±0.1	49.5±0.1
$48 \leq L < 79$	35	45.9±0.1	42.9±0.1	48.9±0.1
$79 \leq L < 293$	35	45.6±0.0	40.7±0.0	50.5±0.0
ALL	140	45.3±0.1	41.7±0.1	48.9±0.1

TABLE X: Metrics of our model on the MeViS dataset under different random seed settings.

Random Seed	val^u			val		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
0	51.4	46.5	56.3	45.4	41.8	49.0
1	51.6	46.7	56.5	45.2	41.6	48.8
2	51.7	46.8	56.6	45.3	41.7	48.9
3	51.4	46.5	56.2	45.4	41.7	49.0
4	51.7	46.8	56.5	45.3	41.7	49.0
5	51.6	46.7	56.5	45.4	41.8	49.0
6	51.5	46.6	56.3	45.4	41.7	49.0
7	51.6	46.7	56.5	45.3	41.6	48.9
8	51.5	46.6	56.4	45.4	41.8	49.0

Fig. 6: **Failure Case of LTCA on MeViS.** Frames where the target of the textual description appears are marked in blue. The segmentation results are shown with a red mask. The objects within the yellow boxes represent the ground-truth targets.

video-object generator to generate masks across frames, the number of parameters in our model is slightly lower than that of LMPM. Our method exhibits a computational complexity similar to that of LMPM, as both our proposed LTCA and the shifted window attention mechanism utilized by LMPM maintain a linear complexity in terms of GLOPs.

Visualizations. Fig. 4 shows qualitative results of both our method and LMPM. While LMPM is prone to segmenting wrong objects, ours in contrast is more accurate in localizing the target objects according to the text. This improvement can be attributed to our proposed LTCA which provides superior long-range temporal context modeling for videos. In Fig. 5, we visualize the results compared with MUTR on Ref-YouTube-VOS benchmark. From Fig. 5, our method can successfully track and segment the referred instance even in challenging situations.

Robustness of LTCA's Random Attention. Random Atten-

tion in LTCA allows object embeddings of each frame to attend to the embeddings of other frames randomly, which introduces randomness into the inference stage. The results presented in Tables I, II, and III validate the robustness of our method across datasets of varying complexity. We conducted the following ablations to further verify the robustness of our proposed random attention. Firstly, we validate LTCA on MeViS under different random seed settings in Table X, which demonstrates that our proposed LTCA is robust. We further validate the stability of random attention under different video lengths (Table IX). Based on the results presented above, we note that all experiments exhibit quite small standard deviations (approximately 0.1%), validating the stability of random attention under different video lengths and video complexities.

D. Failure Case Study

We observe that LTCA may not perform well in the following two situations: (1) dealing with long videos containing scarce frames of target object (as shown in Fig. 6 (a, b)); (2) inferring relative spatial positions (as shown in Fig. 6 (c, d)). In future, we may pre-estimate the text-related frames before aggregating temporal context information and utilize Large Language Models (LLMs) to enhance the understanding of relative spatial positions to deal with the above two limitations respectively.

V. Conclusion

In this paper, we propose a novel attention mechanism, LTCA, designed to capture the long-range temporal context for RVOS. LTCA employs a stack of sparse local attention layers, including dilated window attention and random attention, to strike a balance between locality and globality. It also directly aggregates global information through specifically designed global queries. Our simplified framework achieves new state-of-the-art performance on four RVOS benchmarks compared to previous solutions. We hope that our proposed LTCA and simplified framework may inspire future studies on long video-related tasks.

References

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

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

[3] P. Li, Y. Zhang, L. Yuan, J. Zhao, X. Xu, and X. Zhang, "Adversarial attacks on video object segmentation with hard region discovery," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[4] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, and J. Feng, "Towards real world human parsing: Multiple-human parsing in the wild," *arXiv preprint arXiv:1705.07206*, vol. 3, 2017.

[5] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 792–800.

[6] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan, "Self-supervised neural aggregation networks for human parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 7–15.

[7] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sensing*, vol. 13, no. 21, p. 4441, 2021.

[8] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[9] J. Wang and G. Kang, "Learn to rectify the bias of clip for unsupervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4102–4112.

[10] C. Yan, H. Wang, J. Liu, X. Jiang, Y. Hu, X. Tang, G. Kang, and E. Gavves, "Piclick: Picking the desired mask from multiple candidates in click-based interactive segmentation," *Neurocomputing*, vol. 599, p. 128083, 2024.

[11] M. Wang, K. Chen, L. Su, C. Yan, S. Xu, H. Zhang, P. Yuan, X. Jiang, and B. Zhang, "Rsbuilding: Towards general remote sensing image building extraction and change detection with foundation model," *arXiv preprint arXiv:2403.07564*, 2024.

[12] J. Wang and G. Kang, "Reclip++: Learn to rectify the bias of clip for unsupervised semantic segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2408.06747

[13] K. Chen, X. Jiang, H. Wang, C. Yan, Y. Gao, X. Tang, Y. Hu, and W. Xie, "Ov-dar: Open-vocabulary object detection and attributes recognition," *International Journal of Computer Vision*, pp. 1–23, 2024.

[14] Y. Wang, W. Zhuo, Y. Li, Z. Wang, Q. Ju, and W. Zhu, "Fully self-supervised learning for semantic segmentation," *arXiv preprint arXiv:2202.11981*, 2022.

[15] W. Zhu, S. Nicolau, L. Soler, A. Hostettler, J. Marescaux, and Y. Rémond, "Fast segmentation of abdominal wall: Application to sliding effect removal for non-rigid registration," in *International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging*. Springer, 2012, pp. 198–207.

[16] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring multi-object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 633–14 642.

[17] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," *arXiv preprint arXiv:2309.04379*, 2023.

[18] Z. Yang, Y. Tang, L. Bertinetto, H. Zhao, and P. H. Torr, "Hierarchical interaction network for video object segmentation from referring expressions," in *British Machine Vision Association*, 2021.

[19] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–37, 2013.

[20] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with referring expressions," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[21] J. Mei, A. Piergiovanni, J.-N. Hwang, and W. Li, "Slvp: Self-supervised language-video pre-training for referring video object segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 507–517.

[22] S. Gross, B. Krenn, and M. Scheutz, "Multi-modal referring expressions in human-human task descriptions and their implications for human-robot interaction," *Interaction Studies*, vol. 17, no. 2, pp. 180–210, 2016.

[23] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell *et al.*, "Grounding spatial relations for human-robot interaction," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1640–1647.

[24] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy, "Mevis: A large-scale benchmark for video segmentation with motion expressions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2694–2703.

[25] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3300–3315, 2021.

[26] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph lstm for group activity recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 636–647, 2019.

[27] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1110–1118, 2019.

[28] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Higcin: Hierarchical graph-based cross inference network for group activity recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 6, pp. 6955–6968, 2020.

[29] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7559–7576, 2022.

[30] X. Shu, L. Zhang, Y. Sun, and J. Tang, "Host–parasite: Graph lstm-in-lstm for group activity recognition," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 663–674, 2020.

[31] S. Seo, J.-Y. Lee, and B. Han, "Urvos: Unified referring video object segmentation network with a large-scale benchmark," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 208–223.

[32] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with language referring expressions," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*. Springer, 2019, pp. 123–141.

[33] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek, "Actor and action video segmentation from a sentence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5958–5966.

[34] D.-A. Huang, Z. Yu, and A. Anandkumar, "Minvis: A minimal video instance segmentation framework without video-based training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 265–31 277, 2022.

[35] Z. Jiang, Z. Gu, J. Peng, H. Zhou, L. Liu, Y. Wang, Y. Tai, C. Wang, and L. Zhang, "Stc: spatio-temporal contrastive learning for video instance segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 539–556.

[36] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 588–605.

[37] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5188–5197.

[38] S. Yang, Y. Fang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Crossover learning for fast online video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8043–8052.

[39] L. Yan, Q. Wang, S. Ma, J. Wang, and C. Yu, "Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 393–406, 2022.

[40] H. Fang, T. Zhang, X. Zhou, and X. Zhang, "Learning better video query with sam for video instance segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[41] H. Wang, C. Yan, K. Chen, X. Jiang, X. Tang, Y. Hu, G. Kang, W. Xie, and E. Gavves, "Ov-vis: Open-vocabulary video instance segmentation," *International Journal of Computer Vision*, pp. 1–18, 2024.

[42] T. Zhang, X. Tian, Y. Wu, S. Ji, X. Wang, Y. Zhang, and P. Wan, "Dvis: Decoupled video instance segmentation framework," *arXiv preprint arXiv:2306.03413*, 2023.

[43] T. Zhang, X. Tian, Y. Zhou, S. Ji, X. Wang, X. Tao, Y. Zhang, P. Wan, Z. Wang, and Y. Wu, "Dvis++: Improved decoupled framework for universal video segmentation," *arXiv preprint arXiv:2312.13305*, 2023.

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

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[46] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Video instance segmentation using inter-frame communication transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 352–13 363, 2021.

[47] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "Seqformer: Sequential transformer for video instance segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 553–569.

[48] M. Heo, S. Hwang, S. W. Oh, J.-Y. Lee, and S. J. Kim, "Vita: Video instance segmentation via object token association," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 109–23 120, 2022.

[49] C. Shang, H. Li, H. Qiu, Q. Wu, F. Meng, T. Zhao, and K. N. Ngan, "Cross-modal recurrent semantic comprehension for referring image segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3229–3242, 2022.

[50] H. Li, M. Sun, J. Xiao, E. G. Lim, and Y. Zhao, "Fully and weakly supervised referring expression segmentation with end-to-end learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5999–6012, 2023.

[51] S. Wu, S. Jin, W. Zhang, L. Xu, W. Liu, W. Li, and C. C. Loy, "F-lmm: Grounding frozen large multimodal models," *arXiv preprint arXiv:2406.05821*, 2024.

[52] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 108–124.

[53] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5745–5753.

[54] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7454–7463.

[55] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7436–7456, 2021.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[57] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 38–54.

[58] G. Feng, Z. Hu, L. Zhang, J. Sun, and H. Lu, "Bidirectional relationship inferring network for referring image localization and segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[59] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 155–18 165.

[60] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," *arXiv preprint arXiv:2303.02153*, 2023.

[61] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," *arXiv preprint arXiv:2308.00692*, 2023.

[62] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin, "Pixellm: Pixel reasoning with large multimodal model," *arXiv preprint arXiv:2312.02228*, 2023.

[63] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[64] R. Fang, S. Yan, Z. Huang, J. Zhou, H. Tian, J. Dai, and H. Li, "Instructseq: Unifying vision tasks with instruction-conditioned multi-modal sequence generation," *arXiv preprint arXiv:2311.18835*, 2023.

[65] T.-H. Wu, G. Biamby, D. Chan, L. Dunlap, R. Gupta, X. Wang, J. E. Gonzalez, and T. Darrell, "See, say, and segment: Teaching lmms to overcome false premises," *arXiv preprint arXiv:2312.08366*, 2023.

[66] W. Chen, D. Hong, Y. Qi, Z. Han, S. Wang, L. Qing, Q. Huang, and G. Li, "Multi-attention network for compressed video referring object segmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4416–4425.

[67] X. Li, J. Wang, X. Xu, X. Li, Y. Lu, and B. Raj, "Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency," *arXiv preprint arXiv:2207.01203*, 2022.

[68] J. Tang, G. Zheng, and S. Yang, "Temporal collection and distribution for referring video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 466–15 476.

[69] M. Gao, J. Yang, J. Han, K. Lu, F. Zheng, and G. Montana, "Decoupling multimodal transformers for referring video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[70] B. Miao, M. Bennamoun, Y. Gao, M. Shah, and A. Mian, "Temporally consistent referring video object segmentation with hybrid memory," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[71] M. Sun, J. Xiao, E. G. Lim, C. Zhao, and Y. Zhao, "Unified multi-modality video object segmentation using reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[72] Z. Luo, Y. Xiao, Y. Liu, S. Li, Y. Wang, Y. Tang, X. Li, and Y. Yang, "Soc: Semantic-assisted object cluster for referring video object segmentation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[73] S. Yan, R. Zhang, Z. Guo, W. Chen, W. Zhang, H. Li, Y. Qiao, Z. He, and P. Gao, "Referred by multi-modality: A unified temporal transformer for video object segmentation," *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[74] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[75] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.
[76] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[77] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.
[78] Z. Ding, T. Hui, J. Huang, X. Wei, J. Han, and S. Liu, "Language-bridged spatial-temporal interaction for referring video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4964–4973.
[79] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vlt: Vision-language transformer and query generation for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[80] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.
[81] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.
[82] D. Wu, X. Dong, L. Shao, and J. Shen, "Multi-level representation learning with semantic alignment for referring video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4996–5005.
[83] D. Wu, T. Wang, Y. Zhang, X. Zhang, and J. Shen, "Onlinerefer: A simple online baseline for referring video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2761–2770.
[84] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 502–10 511.
[85] D. Li, R. Li, L. Wang, Y. Wang, J. Qi, L. Zhang, T. Liu, Q. Xu, and H. Lu, "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1297–1305.
[86] J. Wu, Y. Jiang, B. Yan, H. Lu, Z. Yuan, and P. Luo, "Segment every reference object in spatial and temporal spaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2538–2550.
[87] H. Wang, C. Deng, J. Yan, and D. Tao, "Asymmetric cross-guided attention network for actor and action video segmentation from natural language query," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3939–3948.
[88] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4761–4775, 2021.