# Language-Aware Spatial-Temporal Collaboration for Referring Video Segmentation

Tianrui Hui, Si Liu, Zihan Ding, Shaofei Huang, Guanbin Li, Wenguan Wang, Luoqi Liu, and Jizhong Han

***Abstract*—Given a natural language referring expression, the goal of referring video segmentation task is to predict the segmentation mask of the referred object in the video. Previous methods only adopt 3D CNNs upon the video clip as a single encoder to extract a mixed spatio-temporal feature for the target frame. Though 3D convolutions are able to recognize which object is performing the described actions, they still introduce misaligned spatial information from adjacent frames, which inevitably confuses features of the target frame and leads to inaccurate segmentation. To tackle this issue, we propose a language-aware spatial-temporal collaboration framework that contains a 3D temporal encoder upon the video clip to recognize the described actions, and a 2D spatial encoder upon the target frame to provide undisturbed spatial features of the referred object. For multimodal features extraction, we propose a Cross-Modal Adaptive Modulation (CMAM) module and its improved version CMAM+ to conduct adaptive cross-modal interaction in the encoders with spatial- or temporal-relevant language features which are also updated progressively to enrich linguistic global context. In addition, we also propose a Language-Aware Semantic Propagation (LASP) module in the decoder to propagate semantic information from deep stages to the shallow stages with language-aware sampling and assignment, which is able to highlight language-compatible foreground visual features and suppress language-incompatible background visual features for better facilitating the spatial-temporal collaboration. Extensive experiments on four popular referring video segmentation benchmarks demonstrate the superiority of our method over the previous state-of-the-art methods.**

***Index Terms*—Referring video segmentation, spatial-temporal collaboration, cross-modal adaptive modulation, language-aware semantic propagation.**

## I. Introduction

REFERRING video segmentation (RVS) is an emerging task that aims to segment the foreground object in a video described by a natural language referring expression. Compared to video semantic segmentation [2], [3], [4] and semi-supervised video object segmentation [5], [6], [7], RVS is neither restricted by the predefined set of semantic categories nor requires laborious mask annotations in the first frame, thus endowing the RVS model with more flexibility due to the free-form language expressions. Involving both computer vision and natural language processing, RVS task opens up a wide range of potential applications such as human-robot interaction [8], language-driven video editing [9] and intelligent surveillance video processing [10].

In Fig. 1, we present a video clip example where the target frame is in the middle (only 3 frames are shown for brevity) and a referring expression “a white and brown cat is jumping backward”. The goal of RVS is to predict the pixel-level mask of the referred cat on the target frame. Since the prediction relies on the context information of the whole video clip, we claim that both temporal motion modeling over the multiple frames and spatial appearance modeling over the target frame are crucial for tackling the RVS task. On the one hand, spatial modeling alone is insufficient to identify the correct cat by exploiting only appearance clues due to the existence of two white and brown cats in the target frame. Instead, the model inclines to generate plausible but false-positive predictions on other cats. Therefore, incorporating information from adjacent frames is required to recognize the described action for distinguishing the jumping cat from the sitting one. As a result, temporal modeling over the multiple frames serves as an indispensable component. On the other hand, as the jumping cat possesses various poses and locations in 3 frames, temporal modeling will inevitably disturb the feature representation of the target frame by aggregating features of the spatially-misaligned pixels from adjacent frames. The correspondence between the target frame and its ground-truth mask is hence obscured to some extent, which also leads to the necessity of spatial modeling over the target frame to compensate for undisturbed and precise spatial features.

However, previous methods [11], [12], [13], [14], [15] perform entangled spatial-temporal modeling over the multiple frames. They first feed the video clip into a temporal encoder

Manuscript received 28 February 2022; revised 24 December 2022; accepted 2 January 2023. Date of publication 10 January 2023; date of current version 5 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62122010 and 61976250, in part by the Fundamental Research Funds for the Central Universities, CAAI-Huawei MindSpore Open Fund, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515020048. Recommended for acceptance by Cees G. M. Snoek. *(Corresponding author: Si Liu.)*

Tianrui Hui, Shaofei Huang, and Jizhong Han are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100045, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Huairou 101408, China (e-mail: huitianrui@gmail.com; nowherespyfly@gmail.com; hanjizhong@iie.ac.cn).

Si Liu and Zihan Ding are with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China, and also with the Hangzhou Innovation Institute, Beihang University, Beijing 100191, China (e-mail: liusi@buaa.edu.cn; dingzihan737@gmail.com).

Guanbin Li is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510275, China (e-mail: liguanbin@mail.sysu.edu.cn).

Wenguan Wang is with the ReLER Lab, Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: wenguanwang.ai@gmail.com).

Luoqi Liu is with Meitu Inc, Xiamen 361008, China (e-mail: llq5@meitu.com).

Code is available at https://github.com/RyanHTR/CSTM.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPAMI.2023.3235720.

Digital Object Identifier 10.1109/TPAMI.2023.3235720

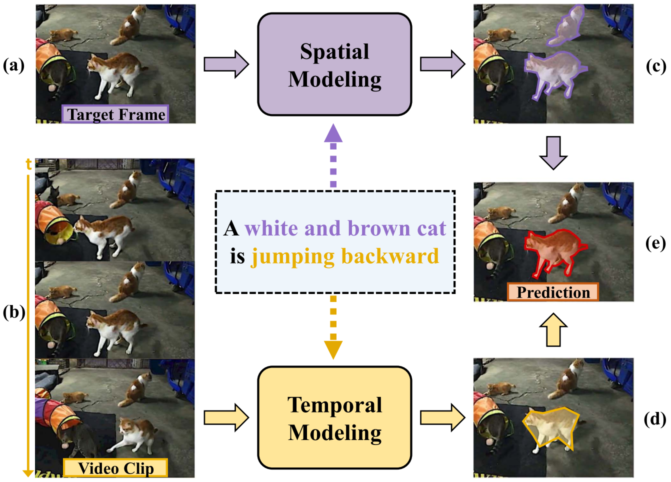


Fig. 1. Illustration of our motivation. (a) The target frame. (b) The input video clip. (c) Spatial modeling alone can generate plausible segmentation but may misidentify other objects due to insufficient action recognition ability. (d) Temporal modeling alone can distinguish the correct object which performs the described action but may introduce misaligned spatial features into the target frame, yielding inaccurate segmentation. (e) Through language-aware spatial-temporal collaboration, the correct object in the target frame can be well segmented.

(3D CNN [16]) to extract multi-frame video features, then pool out the temporal dimension to obtain a mixed spatio-temporal feature of the target frame. According to the above discussion, the feature of the target frame will be confused by mixing multi-frame spatial information, yielding inaccurate segmentation. To tackle this limitation, we propose a *language-aware spatial-temporal collaboration* framework to conduct spatial modeling over the target frame and temporal modeling over the multiple frames respectively with two independent multi-modal encoders. For the temporal encoder, we adopt a 3D CNN [16] to identify the object performing the described actions, which serves as the coarse localization of the correct object. For the spatial encoder, we adopt a 2D CNN [17] to provide undisturbed and precise spatial feature of the target frame, which can be regarded as the fine segmentation of the referred object.

In addition, the referring expression also contains both spatial-relevant information (appearance words, e.g., “white and brown”) and temporal-relevant information (action words, e.g., “jumping backward”). To accomplish adaptive cross-modal interaction in the two encoders, features of spatial-relevant words can play a bigger role when interacting with the visual feature from the spatial encoder, and vice versa for the temporal encoder. Therefore, we propose a Cross-Modal Adaptive Modulation (CMAM) module utilizing cross-modal attention to dynamically recombine spatial- or temporal-relevant language features, by which the corresponding visual features are adaptively modulated. The CMAM module is densely inserted into each stage of the two encoders, thus enabling visual features to hierarchically interact with language features and highlighting regions of the referred object. For CMAM in each stage, we further supplement the recombined language feature to original words features with self-attention, which enriches the global context of words features with cross-modal spatial or temporal preference. The enhanced words features serve as the input of the next stage, which forms a progressive language update path through the visual encoders for more effective cross-modal interaction. We denote this improved module as CMAM+ which bears obvious superiority over the original CMAM module.

In the decoder, spatial and temporal visual features from two encoders are fused and propagated between adjacent stages for comprehensive spatial-temporal collaboration. Concretely, we also propose a Language-Aware Semantic Propagation (LASP) module that first samples foreground pixel features most compatible with the language and background pixel features most incompatible with the language in the deep stage of the decoder. Then, these spatial-temporal pixel features are propagated to the corresponding positions in the adjacent shallow stage to selectively highlight language-compatible foreground features and suppress language-incompatible background features. By this means, our LASP can effectively aggregate high-level semantic information in deep stages and low-level local details in shallow stages through language sampling and propagation. The aggregated spatial-temporal features establish a comprehensive collaboration process that identifies the referred object and refines the mask prediction in a stage-wise manner.

In summary, this paper has the following contributions:

- We propose a language-aware spatial-temporal collaboration framework that consists of a temporal encoder to recognize the described action and a spatial encoder to provide undisturbed and precise spatial features of the referred object based on language clues. Collaborative spatial-temporal modeling can help the model better identify and segment the referred object.
- In the encoder, we also propose a Cross-Modal Adaptive Modulation (CMAM) module and its improved version CMAM+ to conduct adaptive cross-modal interaction with spatial- or temporal-relevant language features which are also updated in each stage.
- In the decoder, a Language-Aware Semantic Propagation (LASP) module is further proposed to highlight language-compatible foreground visual features and suppress language-incompatible background visual features by cross-stage feature sampling and semantic propagation, which further facilitates the spatial-temporal collaboration.
- Extensive experiments on four popular referring video segmentation benchmarks, i.e., A2D Sentences [11], J-HMDB Sentences [11], Refer-YouTube-VOS [18] and Refer-DAVIS [19], demonstrate that our method outperforms previous state-of-the-art methods.

This paper is built upon our conference version [1] and significantly extends it in several aspects. First, we extend the original CMAM module as CMAM+ where the global context of recombined language feature is supplemented to initial words features, forming a language update path through the encoders to make language more adaptive to visual features in different stages. Second, we newly propose a LASP module in the decoder to highlight foreground features and suppress background features via language-aware feature sampling and semantic propagation from deep stages to shallow ones, by which the spatial-temporal collaboration is further facilitated with aggregated multi-level contexts. Third, we make our model end-to-end trainable instead of training encoders and decoder in two stages for better performance and training efficiency. Fourth, we add considerable new experiments including ablation study, qualitative analysis, and comparison results on two more popular benchmarks. Our extended model achieves large performance gains (5.2% mAP on A2D Sentences and 5.7% mAP on J-HMDB Sentences) over our conference version.

## II. Related Work

### A. Referring Image Segmentation

Similar to RVS, referring image segmentation (RIS) aims to segment the referred object in a static image. Early works [20], [21], [22], [23] follow a concatenation-and-convolution scheme to directly fuse visual and linguistic features with recurrent refinement, dynamic filtering, or multi-scale context. Cross-modal attention [24], [25], [26] combined with word semantic classification [27] or linguistic structure analysis [28] further improves the segmentation performance. MCN [29] proposes a multi-task learning framework where referring localization and segmentation can mutually refine each other by explicit loss constraints. Position prior [30] and bottom-up visual reasoning [31] are also exploited to gradually identify the regions of referred objects. Recently, Transformer [32] has shown a notable ability to capture long-range dependencies for feature extraction in both language and vision communities. VLT [33] follows the Transformer encoder-decoder framework of DETR [34] where different combinations of language are directly utilized as queries to find the most responsive regions on the image, which achieves notable performance. LAVT [35] conducts visual and linguistic feature fusion early in the intermediate stages of the Vision Transformer backbone so that better cross-modal alignment can be achieved progressively through the visual encoder. CRIS [36] proposes to transfer the rich vision-language prior knowledge from the CLIP model [37] to the RIS task by designing text-to-pixel contrastive learning for text-to-pixel alignment. Furthermore, DenseCLIP [38] transfers the pre-trained knowledge of CLIP to more general dense prediction tasks (e.g, object detection, instance segmentation, etc) by converting image-text matching to pixel-text matching and prompting the language model using visual contexts. In this paper, we focus on the referring segmentation task on video data where temporal information involving multiple frames is essential, and we seek the collaboration of spatial-temporal modeling to better align video and language modalities.

### B. Referring Video Segmentation

The RVS task is first proposed in [11] where referring expression annotations are provided based on the A2D dataset [39] containing pixel-level labels of both actors and actions. Gavrilyuk et al. [11] proposes to use dynamic filters in the decoder to generate multi-scale cross-modal response maps for mask refinement. Afterwards, Wang et al. [14] further incorporate dynamic filters with deformable convolutions [40] to capture the shape and geometric variations of referred objects. ACGA [12] exploits asymmetric cross-attention mechanism to perform spatial information exchange between visual and linguistic features for cross-modal matching with enhanced multimodal context. In [13], capsule networks [41] are introduced to encode video and language features with dynamic routing, which explores a different multimodal representation other than naive convolutions. In addition to feature extracting, polar positional encoding [15] is also devised to encode richer positional information than general absolute or relative positional encodings, yielding better localization results of the referred objects. CMPC-V [42] performs entity, relation, and action reasoning on joint visual and linguistic features with word analysis. CMSA [43] extends the cross-modal self-attention to multiple frames for temporal context extraction.

Since the common video object segmentation datasets (e.g., DAVIS [44] and YouTube-VOS [45]) also contain pixel-level labeling of multiple instances in videos, some works extend them with referring expression annotations and propose Refer-DAVIS [19] and Refer-YouTube-VOS [18] benchmarks for dense RVS prediction. In [19], a two-stage grounding-then-segmenting pipeline is applied on the video input while Seo et al. [18] utilize spatial-temporal memory attention to propagate features and predictions of previous frames to the current target frame. HINet [46] proposes to hierarchically fuse language features with visual features from different levels of the visual encoder so that different concepts in the language can be extracted. Recently, ReferFormer [47] follows the paradigms of DETR [34] and VisTR [48] where language features are adopted as queries to attend to different objects in the video frames by dense attention operations in the Transformer encoder and decoder. MTTR [49] leverages a similar framework as ReferFormer with instance queries and utilizes a multimodal Transformer encoder to fuse vision and language features. Different from these works, our method proposes a spatial-temporal collaboration framework that exploits language to adaptively modulate the visual features in the encoders and highlight or suppress language-compatible or -incompatible visual features in the decoder.

### C. Semi-Supervised Video Object Segmentation

Given the manually-labeled masks of objects in the first frame, the semi-supervised video object segmentation task aims to segment these objects in the entire video sequence. The main methodology of many works [7], [50], [51], [52], [53] is to perform pixel-level feature similarity matching. FEELVOS [50] exploits the information of the first frame and previous frame to generate the template and the network is guided by the feature matching result. EGMN [7] organizes the network as a fully-connected graph where frames are stored as nodes and cross-frame correlations are captured by edges. AOT [53] utilizes Transformers to match and decode multiple objects simultaneously in the same embedding space. Some works [50], [54], [55] also focus on suppressing the background distraction. CFBI [55] employs multiple windows to perform local matching to exclude background distractors and capture object motions. For our RVS task, the annotations in the first frame are replaced

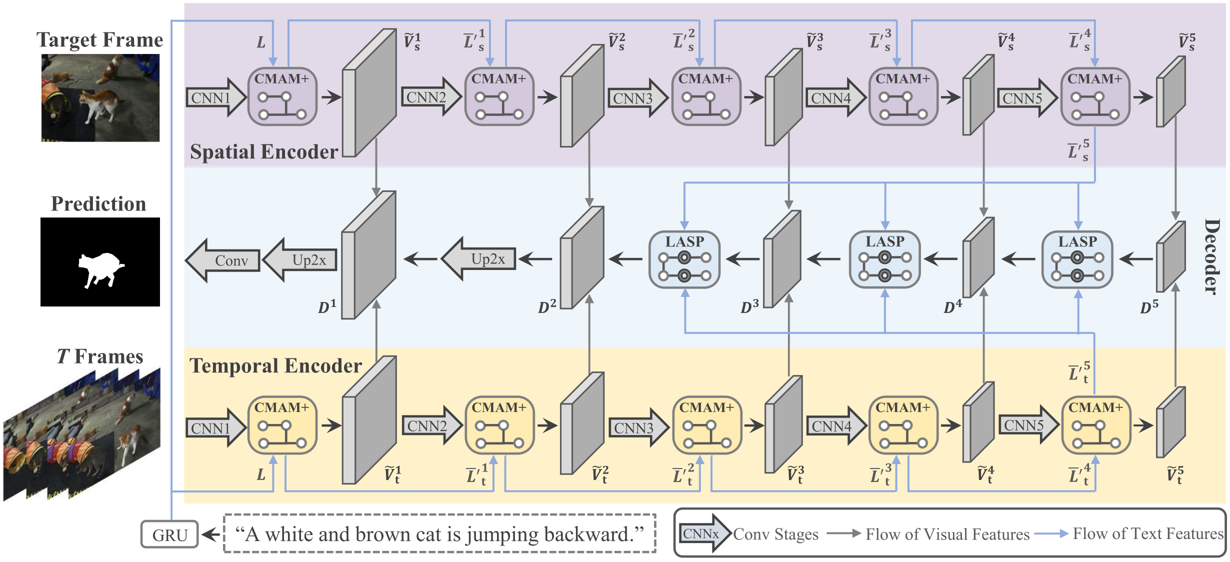

Target Frame
Prediction
T Frames
Spatial Encoder
Temporal Encoder
Decoder
CNN1
CNN2
CNN3
CNN4
CNN5
CMAM+
LASP
Conv
Up2x
L
$\tilde{V}_s^1$ $\tilde{V}_s^2$ $\tilde{V}_s^3$ $\tilde{V}_s^4$ $\tilde{V}_s^5$
$\bar{L}'^1_s$ $\bar{L}'^2_s$ $\bar{L}'^3_s$ $\bar{L}'^4_s$ $\bar{L}'^5_s$
$D^1$ $D^2$ $D^3$ $D^4$ $D^5$
$\tilde{V}_t^1$ $\tilde{V}_t^2$ $\tilde{V}_t^3$ $\tilde{V}_t^4$ $\tilde{V}_t^5$
$\bar{L}'^1_t$ $\bar{L}'^2_t$ $\bar{L}'^3_t$ $\bar{L}'^4_t$ $\bar{L}'^5_t$
GRU
"A white and brown cat is jumping backward."
CNNx Conv Stages
Flow of Visual Features
Flow of Text Features

Fig. 2. Overall architecture of our method. Spatial and temporal encoders extract features of the target frame and the video clip respectively, aided by CMAM+ which dynamically interacts with multimodal features in each stage. LASP is also densely applied in adjacent stages of the decoder (except for the shallowest stage) to highlight language-compatible foreground features and suppress language-incompatible background features.

by referring expressions so that both spatial-temporal collaboration and vision-language alignment are important.

### D. Spatial-Temporal Modeling

In order to solve video-related tasks, spatial-temporal modeling [56], [57], [58] is an essential component. A direct way of conducting spatial-temporal modeling is applying 3D CNNs (e.g., C3D [59] and I3D [16]) to extract multi-frame features. (2+1)D ConvNets [60], [61] are further proposed to decompose 3D convolution for reducing its computational budget. SlowFast [62] proposes a slow path along with a fast path to capture spatial and motion information respectively with different sampling rates. TSM [63] proposes a lightweight temporal shift module where a portion of feature channels on the time dimension is shifted to approximately model temporal information, and this operation can be generalized as a 1D temporal convolution [64]. In this paper, our model shares the same spirit with SlowFast where the collaboration of spatial and temporal modeling with language clues is proposed to jointly extract the precise spatial information and inter-frame temporal context.

## III. Method

Fig. 2 illustrates the overall architecture of our proposed method. The input target frame, video clip and referring expression are processed by visual and linguistic encoders respectively (i.e., 2D CNNs [17], 3D CNNs [16] and GRU [65]). For visual input, we adopt a spatial encoder and a temporal encoder to collaboratively extract spatial and temporal visual features respectively. In each stage of the spatial and temporal visual encoders, we apply our improved Cross-Modal Adaptive Modulation module (CMAM+) to dynamically recombine language features to modulate spatial- or temporal-relevant visual features for adaptive cross-modal interaction. The recombined language feature is further supplemented to original words features to supplement global context for progressive language update. Then in the decoder, we proposed a Language-Aware Semantic Propagation (LASP) module which exploits cross-stage feature sampling and semantic propagation to highlight language-compatible foreground visual features and suppress language-incompatible background visual features, thus further facilitating the spatial-temporal collaboration. After stagewise refinement and upsampling, our decoder finally outputs a feature map of the same size as the input target frame to produce the segmentation mask.

### A. Visual and Linguistic Encoders

Given a video clip with $T$ frames, we adopt ResNet-50 [17] as the spatial encoder to process the annotated target frame $\boldsymbol{F}_u \in \mathbb{R}^{H^0\times W^0\times 3}$ in the middle of the clip, and I3D [16] as the temporal encoder to process the whole video clip $\boldsymbol{F} \in \mathbb{R}^{T\times H^0\times W^0\times 3}$ respectively. We denote visual features from the $i$th stage ($i \in [1, 5]$) of the spatial and temporal encoders as $\boldsymbol{V}_{\mathrm{s}}^i \in \mathbb{R}^{H^i\times W^i\times C_{\mathrm{v}}^i}$ and $\boldsymbol{V}_{\mathrm{t}}^i \in \mathbb{R}^{T^i\times H^i\times W^i\times C_{\mathrm{v}}^i}$ respectively, where $H^i = \frac{H^0}{2^i}$, $W^i = \frac{W^0}{2^i}$, and $C_{\mathrm{v}}^i$ are the height, width and channel number of the $i$th visual feature. The reduction of $T^i$ follows the protocol of the I3D network. An 8-dimensional coordinate feature encoding relative position information of each pixel is also adopted following [12]. Since the coordinate feature is appended to visual features of all stages, we omit its notation for simplicity. For the input referring expression containing $N$ words, we first embed each word into a vector using GloVe embeddings [66] and then extract the linguistic feature $\boldsymbol{L} \in \mathbb{R}^{N\times C_{\mathrm{l}}}$ by GRU [65], where $C_{\mathrm{l}}$ denotes the channel number.

### B. Cross-Modal Adaptive Modulation

The goal of our CMAM module is to form an effective interaction between visual and linguistic features by adaptive visual

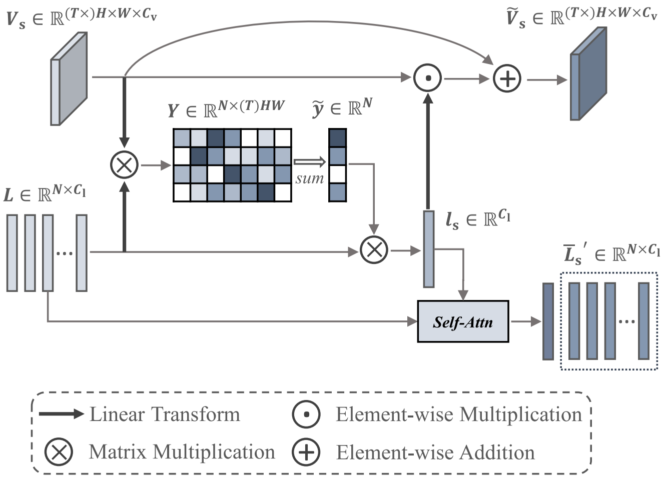

$V_s \in \mathbb{R}^{(T\times)H\times W\times C_v}$
$\tilde{V}_s \in \mathbb{R}^{(T\times)H\times W\times C_v}$
$Y \in \mathbb{R}^{N\times(T)HW}$
$\tilde{y} \in \mathbb{R}^N$
sum
$L \in \mathbb{R}^{N\times C_l}$
$l_s \in \mathbb{R}^{C_l}$
$\bar{L}_s{}' \in \mathbb{R}^{N\times C_l}$
Self-Attn
Linear Transform
Element-wise Multiplication
Matrix Multiplication
Element-wise Addition

Fig. 3. Illustration of CMAM+ module. The linguistic feature is dynamically recombined based on the relevance with visual features and further enriched by global context with spatial or temporal preference. The situation of the temporal encoder is denoted in parentheses, which is similar to the spatial encoder.

modulation with linguistic clues, thus highlighting matched visual regions. We insert the CMAM module into each stage of the spatial and temporal encoders so that the cross-modal interaction can be densely refined. To clearly elaborate how CMAM works, we take the $i$th stage of our spatial encoder as an example and omit the superscript $i$ for ease of presentation. As illustrated in Fig. 3, given the visual feature $\boldsymbol{V}_{\mathrm{s}} \in \mathbb{R}^{H\times W\times C_{\mathrm{v}}}$ of the target frame and the linguistic feature $\boldsymbol{L} \in \mathbb{R}^{N\times C_{\mathrm{l}}}$ of the sentence, we first perform cross-modal attention between $\boldsymbol{V}_{\mathrm{s}}$ and $\boldsymbol{L}$ to obtain an attention map $\boldsymbol{Y} \in \mathbb{R}^{N\times HW}$ which computes the feature relevance between the target frame and each word. In detail, $\boldsymbol{V}_{\mathrm{s}}$ and $\boldsymbol{L}$ are first projected to the same subspace by linear layers (reshaping operations are omitted):

$$\boldsymbol{V}_{\mathrm{s}}' = \boldsymbol{V}_{\mathrm{s}}\boldsymbol{W}_1, \boldsymbol{L}_{\mathrm{s}}' = \boldsymbol{L}\boldsymbol{W}_2, \tag{1}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{C_{\mathrm{v}}\times C_{\mathrm{m}}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C_{\mathrm{l}}\times C_{\mathrm{m}}}$ are projection parameters, $\boldsymbol{V}_{\mathrm{s}}' \in \mathbb{R}^{H\times W\times C_{\mathrm{m}}}$, $\boldsymbol{L}_{\mathrm{s}}' \in \mathbb{R}^{N\times C_{\mathrm{m}}}$ are projected features. $\boldsymbol{V}_{\mathrm{s}}'$ is then reshaped to $\mathbb{R}^{HW\times C_{\mathrm{m}}}$ to match the matrix dimensions. We further conduct matrix multiplication between $\boldsymbol{V}_{\mathrm{s}}'$ and $\boldsymbol{L}_{\mathrm{s}}'$ to produce the attention map $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \boldsymbol{L}_{\mathrm{s}}'\boldsymbol{V}_{\mathrm{s}}'^{\mathrm{T}}. \tag{2}$$

Here $\boldsymbol{Y} \in \mathbb{R}^{N\times HW}$ calculates the feature relevance between each word and each pixel on the spatial visual feature map. Then, all the values on the $HW$ dimension are summed and normalized as follows:

$$\boldsymbol{y} = \sum_{j=1}^{HW} \boldsymbol{Y}^j,$$
$$\tilde{\boldsymbol{y}}^n = \frac{\exp(\boldsymbol{y}^n/\left\|\boldsymbol{y}\right\|_2)}{\sum_{k=1}^{N}\exp(\boldsymbol{y}^k/\left\|\boldsymbol{y}\right\|_2)}, \tag{3}$$

where $\|\cdot\|_2$ denotes the $\ell_2$-norm of a vector, $\boldsymbol{Y}^j \in \mathbb{R}^N$ is the feature relevance between $N$ words and the $j$th pixel, and $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}^n\}_{n=1}^N \in \mathbb{R}^N$ is the normalized global feature relevance between each word and the whole target frame. Therefore, we can use $\tilde{\boldsymbol{y}}$ to linearly re-combine features of $N$ words to obtain adaptive sentence feature $\boldsymbol{l}_{\mathrm{s}} = \sum_{n=1}^{N}(\tilde{\boldsymbol{y}}^n\boldsymbol{L}^n) \in \mathbb{R}^{C_{\mathrm{l}}}$ which contains more spatial information matched with the spatial visual feature $\boldsymbol{V}_{\mathrm{s}}$ of the target frame.

Afterwards, we adopt a linear layer and the sigmoid function to project $\boldsymbol{l}_{\mathrm{s}}$ to $\mathbb{R}^{C_{\mathrm{v}}}$ dimensions and generate channel-wise modulation weights $\tilde{\boldsymbol{l}}_{\mathrm{s}} \in \mathbb{R}^{C_{\mathrm{v}}}$:

$$\tilde{\boldsymbol{l}}_{\mathrm{s}} = \sigma(\boldsymbol{l}_{\mathrm{s}}\boldsymbol{W}_3), \tag{4}$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{C_{\mathrm{l}}\times C_{\mathrm{v}}}$ is the projection parameter and $\sigma$ denotes sigmoid function. Then, $\tilde{\boldsymbol{l}}_{\mathrm{s}}$ is multiplied with feature of the target frame $\boldsymbol{V}_{\mathrm{s}}$ to highlight sentence-relevant visual feature channels, which shares similar spirits with SENet [67]. The modulated feature is added to the original $\boldsymbol{V}_{\mathrm{s}}$ to ease optimization:

$$\tilde{\boldsymbol{V}}_{\mathrm{s}} = \boldsymbol{V}_{\mathrm{s}} + \boldsymbol{V}_{\mathrm{s}} \odot \tilde{\boldsymbol{l}}_{\mathrm{s}}, \tag{5}$$

where $\odot$ denotes elementwise multiplication.

In the original CMAM module, the language inputs for different encoder stages are identical where the linguistic adaptability w.r.t the visual features is restricted to some extent. To adapt visual features containing information of different abstraction levels, we further improve the CMAM module with an additional language update path through the visual encoders. Concretely, we first concatenate the recombined sentence feature $\boldsymbol{l}_{\mathrm{s}} \in \mathbb{R}^{C_{\mathrm{l}}}$ with the original words features $\boldsymbol{L} \in \mathbb{R}^{N\times C_{\mathrm{l}}}$ to form $\hat{\boldsymbol{L}}_{\mathrm{s}} \in \mathbb{R}^{(N+1)\times C_{\mathrm{l}}}$, namely a new language sequence with length of $N+1$. Then, we conduct self-attention on $\hat{\boldsymbol{L}}_{\mathrm{s}}$ to supplement the global linguistic context to original words features with cross-modal spatial or temporal preference as follows:

$$\begin{aligned}\bar{\boldsymbol{L}}_{\mathrm{s}} &= \hat{\boldsymbol{L}}_{\mathrm{s}} + \mathcal{F}_{\mathrm{attn}}(\hat{\boldsymbol{L}}_{\mathrm{s}}, \hat{\boldsymbol{L}}_{\mathrm{s}}, \hat{\boldsymbol{L}}_{\mathrm{s}})\\ &= \hat{\boldsymbol{L}}_{\mathrm{s}} + \mathrm{softmax}\left(\frac{\hat{\boldsymbol{L}}_{\mathrm{s}}\boldsymbol{W}_4(\hat{\boldsymbol{L}}_{\mathrm{s}}\boldsymbol{W}_5)^{\mathrm{T}}}{\sqrt{C_{\mathrm{l}}}}\right)(\hat{\boldsymbol{L}}_{\mathrm{s}}\boldsymbol{W}_6),\end{aligned} \tag{6}$$

where $\mathcal{F}_{\mathrm{attn}}(q,k,v)$ denotes the general attention function based on feature similarity, $\boldsymbol{W}_4$, $\boldsymbol{W}_5$ and $\boldsymbol{W}_6$ are projection parameters for query, key and value. Afterwards, we remove the token corresponding to $\boldsymbol{l}_{\mathrm{s}}$ from $\bar{\boldsymbol{L}}_{\mathrm{s}} \in \mathbb{R}^{(N+1)\times C_{\mathrm{l}}}$ to obtain the enriched features of $N$ words $\bar{\boldsymbol{L}}_{\mathrm{s}}' \in \mathbb{R}^{N\times C_{\mathrm{l}}}$. We denote this improved version of CMAM as CMAM+ where $\bar{\boldsymbol{L}}_{\mathrm{s}}'$ and $\tilde{\boldsymbol{V}}_{\mathrm{s}}$ are its output and serve as the input features of CMAM+ module in the next stage of the spatial visual encoder. For the temporal visual encoder, the same operations are performed on the whole video clip to highlight sentence-relevant temporal visual features in each stage.

### C. Language-Aware Semantic Propagation

After dense cross-modal interaction by CMAM+ in the encoders, spatial and temporal visual features are fused in the decoder to collaboratively generate the segmentation mask of the referred object. Inspired by the spirits of previous works [22], [54], [68], [69], [70] which show the visual features in deep stages contain semantic information and those in shallow stages preserve fine details, integrating multi-stage features can capture the more comprehensive visual context of the referred object

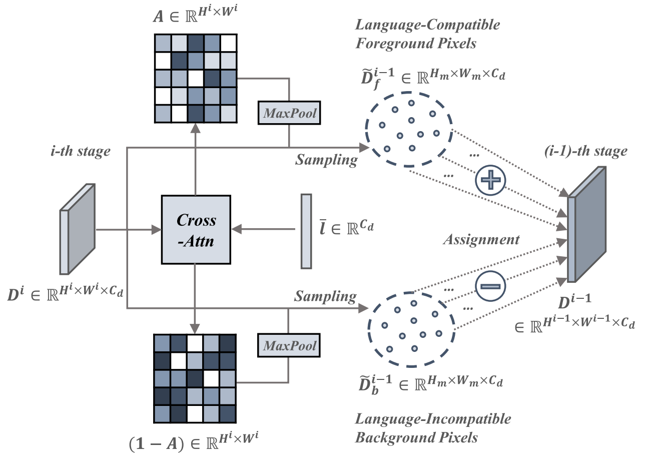

$A \in \mathbb{R}^{H^i \times W^i}$
Language-Compatible Foreground Pixels
$\tilde{D}_f^{i-1} \in \mathbb{R}^{H_m \times W_m \times C_d}$
MaxPool
i-th stage
Sampling
(i-1)-th stage
Cross -Attn
$\bar{l} \in \mathbb{R}^{C_d}$
Assignment
$D^i \in \mathbb{R}^{H^i \times W^i \times C_d}$
Sampling
$D^{i-1} \in \mathbb{R}^{H^{i-1} \times W^{i-1} \times C_d}$
MaxPool
$\tilde{D}_b^{i-1} \in \mathbb{R}^{H_m \times W_m \times C_d}$
Language-Incompatible Background Pixels
$(1 - A) \in \mathbb{R}^{H^i \times W^i}$

Fig. 4. Illustration of LASP module. The relevance score map between linguistic and visual features is exploited to sample language-compatible foreground pixels and language-incompatible background pixels from the deep stage. Foreground and background visual features in the shallow stage are accordingly highlighted and suppressed by the propagated deep semantics.

and refine the mask quality. To this end, we further propose a new Language-Aware Semantic Propagation (LASP) module where foreground pixels most compatible with language and background pixels most incompatible with language are sampled from features of deep stages. Then, the sampled language-aware deep semantics are propagated to the shallow stages to highlight foreground features and suppress background features for accurately segmenting the referred object. The LASP module is inserted into the adjacent stages of the decoder so that the deep semantics are progressively propagated to the shallow stages and recover feature resolution.

As illustrated in Fig. 4, let $\tilde{\boldsymbol{V}}_{\mathrm{s}}^{i}$ and $\tilde{\boldsymbol{V}}_{\mathrm{t}}^{i}$ denote features from the $i$th stage ($i = 3, 4, 5$) of spatial encoder and temporal encoder (target frame is selected) respectively. We first sum them to obtain spatial-temporal visual feature $\boldsymbol{D}^i \in \mathbb{R}^{H^i \times W^i \times C_{\mathrm{d}}}$. Then, the words features updated by CMAM+ modules in the last stages of spatial and temporal encoders, i.e., $\bar{\boldsymbol{L}}_{\mathrm{s}}'$ and $\bar{\boldsymbol{L}}_{\mathrm{t}}'$ (stage indexes are omitted), are averaged to obtain the global sentence feature $\bar{\boldsymbol{l}} \in \mathbb{R}^{C_{\mathrm{d}}}$:

$$\bar{\boldsymbol{l}} = \left[ \sum_{k=1}^{N} \bar{\boldsymbol{L}}_{\mathrm{s}}^{\prime,k}; \sum_{k=1}^{N} \bar{\boldsymbol{L}}_{\mathrm{t}}^{\prime,k} \right] \boldsymbol{W}_7, \tag{7}$$

where $\boldsymbol{W}_7 \in \mathbb{R}^{C_{\mathrm{l}} \times C_{\mathrm{d}}}$ is the parameter of linear layer and $[;]$ denotes feature concatenation. We further perform relevance filtering between $\bar{\boldsymbol{l}}$ and $\boldsymbol{D}^i$ to yield the relevance score map $\boldsymbol{A} \in \mathbb{R}^{1 \times H^i W^i}$ which measures the feature compatibility between pixels and the whole sentence:

$$\boldsymbol{A} = \sigma(\bar{\boldsymbol{l}} \boldsymbol{W}_8 (\boldsymbol{D}^i \boldsymbol{W}_9)^{\mathrm{T}}), \tag{8}$$

where $\boldsymbol{W}_8$ and $\boldsymbol{W}_9 \in \mathbb{R}^{C_{\mathrm{d}} \times C_{\mathrm{d}}}$ are projection parameters and $\sigma$ is sigmoid function. Afterwards, we conduct adaptive max pooling on the reshaped $\boldsymbol{A} \in \mathbb{R}^{H^i \times W^i}$ to obtain a small map with size of $H_{\mathrm{m}} \times W_{\mathrm{m}}$ which selects foreground pixels most compatible with language. Then, the small map is upsampled to the original size of $H^i \times W^i$ to form a soft foreground sampling mask for extracting the language-compatible foreground feature from $\boldsymbol{D}^i$ as follows:

$$\bar{\boldsymbol{D}}^i = \mathcal{F}_{\mathrm{up}}(\mathcal{F}_{\mathrm{maxp}}(\boldsymbol{A})) \odot \boldsymbol{D}^i, \tag{9}$$

where $\mathcal{F}_{\mathrm{maxp}}$ and $\mathcal{F}_{\mathrm{up}}$ denote max pooling and bilinear upsampling. We utilize the indexes of max pooling to sample the most language-compatible foreground pixels from $\bar{\boldsymbol{D}}^i$ and save their features as $\bar{\boldsymbol{D}}_{\mathrm{f}}^i \in \mathbb{R}^{H_{\mathrm{m}} \times W_{\mathrm{m}} \times C_{\mathrm{d}}}$.

The indexes of maximum values are transformed to the corresponding positions on the feature of the previous stage $\boldsymbol{D}^{i-1} \in \mathbb{R}^{H^{i-1} \times W^{i-1} \times C_{\mathrm{d}}}$ so that the foreground features of the $(i-1)$th stage can also be sampled as $\bar{\boldsymbol{D}}_{\mathrm{f}}^{i-1} \in \mathbb{R}^{H_{\mathrm{m}} \times W_{\mathrm{m}} \times C_{\mathrm{d}}}$. Then, we conduct cross-stage attention between $\bar{\boldsymbol{D}}_{\mathrm{f}}^{i}$ and $\bar{\boldsymbol{D}}_{\mathrm{f}}^{i-1}$ to propagate foreground semantics from deep stages to the shallow stages:

$$\tilde{\boldsymbol{D}}_{\mathrm{f}}^{i-1} = \bar{\boldsymbol{D}}_{\mathrm{f}}^{i-1} + \mathcal{F}_{\mathrm{attn}}(\bar{\boldsymbol{D}}_{\mathrm{f}}^{i-1}, \bar{\boldsymbol{D}}_{\mathrm{f}}^{i}, \bar{\boldsymbol{D}}_{\mathrm{f}}^{i}), \tag{10}$$

where $\mathcal{F}_{\mathrm{attn}}(q, k, v)$ denotes the attention function based on feature similarity as mentioned in (6). The obtained $\tilde{\boldsymbol{D}}_{\mathrm{f}}^{i-1} \in \mathbb{R}^{H_{\mathrm{m}} \times W_{\mathrm{m}} \times C_{\mathrm{d}}}$ is assigned to the corresponding positions on $\boldsymbol{D}^{i-1} \in \mathbb{R}^{H^{i-1} \times W^{i-1} \times C_{\mathrm{d}}}$ with the transformed indexes to replace those features, thus highlighting language-compatible foreground visual regions.

In addition, we also exploit the reversed $\boldsymbol{A}$ to yield a soft background sampling mask for extracting the language-incompatible background feature from $\boldsymbol{D}^i$ as follows:

$$\hat{\boldsymbol{D}}^i = \mathcal{F}_{\mathrm{up}}(\mathcal{F}_{\mathrm{maxp}}(1 - \boldsymbol{A})) \odot \boldsymbol{D}^i. \tag{11}$$

Similarly, the max pooling indexes are used to sample the most language-incompatible background pixels from $\hat{\boldsymbol{D}}^i$ and save their features as $\bar{\boldsymbol{D}}_{\mathrm{b}}^i \in \mathbb{R}^{H_{\mathrm{m}} \times W_{\mathrm{m}} \times C_{\mathrm{d}}}$. The corresponding background features on $\boldsymbol{D}^{i-1}$ are also sampled as $\bar{\boldsymbol{D}}_{\mathrm{b}}^{i-1}$ and we perform similar cross-stage attention as (10):

$$\tilde{\boldsymbol{D}}_{\mathrm{b}}^{i-1} = \bar{\boldsymbol{D}}_{\mathrm{b}}^{i-1} - \mathcal{F}_{\mathrm{attn}}(\bar{\boldsymbol{D}}_{\mathrm{b}}^{i-1}, \bar{\boldsymbol{D}}_{\mathrm{b}}^{i}, \bar{\boldsymbol{D}}_{\mathrm{b}}^{i}) \tag{12}$$

where the propagated background semantics are subtracted from $\bar{\boldsymbol{D}}_{\mathrm{b}}^{i-1}$ to suppress language-incompatible background visual regions. The obtained $\tilde{\boldsymbol{D}}_{\mathrm{b}}^{i-1}$ is also assigned to the corresponding positions on $\boldsymbol{D}^{i-1}$ and then fused with previous $\boldsymbol{D}^{i-1}$ assigned by $\tilde{\boldsymbol{D}}_{\mathrm{f}}^{i-1}$ to yield the final output of our LASP on the $(i-1)$th stage. After progressive upsampling, our decoder outputs the visual feature with the same resolution as the input frame and applies convolutions to predict the binary segmentation mask.

## IV. Experiments

### A. Datasets and Evaluation Metrics

We conduct experiments on four popular referring video segmentation benchmarks. Details are presented as follows:

*A2D Sentences [11]:* This dataset is extended from the Actor-Action Dataset [39] (A2D) with 6,655 referring expression annotations in total. It consists of 8 actions categories performed by 7 actors categories with a total number of 3,782 videos collected from YouTube. In each video, 3 to 5 frames are labeled

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE A2D SENTENCES TESTING SET

| Method | Pub. | Overlap | | | | | mAP | IoU | | FPS |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P@0.5 | P@0.6 | P@0.7 | P@0.8 | P@0.9 | 0.5:0.95 | Overall | Mean | |
| Gavrilyuk *et al.* [11] | CVPR18 | 47.5 | 34.7 | 21.1 | 8.0 | 0.2 | 19.8 | 53.6 | 42.1 | - |
| Gavrilyuk *et al.* † [11] | CVPR18 | 50.0 | 37.6 | 23.1 | 9.4 | 0.4 | 21.5 | 55.1 | 42.6 | - |
| ACGA [12] | ICCV19 | 55.7 | 45.9 | 31.9 | 16.0 | 2.0 | 27.4 | 60.1 | 49.0 | 9.5 |
| VT-Capsule [13] | CVPR20 | 52.6 | 45.0 | 34.5 | 20.7 | 3.6 | 30.3 | 56.8 | 46.0 | - |
| CMDy [14] | AAAI20 | 60.7 | 52.5 | 40.5 | 23.5 | 4.5 | 33.3 | 62.3 | 53.1 | 7.9 |
| PRPE [15] | IJCAI20 | 63.4 | 57.9 | 48.3 | 32.2 | 8.3 | 38.8 | 66.1 | 52.9 | 6.3 |
| CMSA [43] | TPAMI21 | 48.7 | 43.1 | 35.8 | 23.1 | 5.2 | - | 61.8 | 43.2 | - |
| CMPC-V [42] | TPAMI21 | 65.5 | 59.2 | 50.6 | 34.2 | 9.8 | 40.4 | 65.3 | 57.3 | 12.2 |
| Ours [1] | CVPR21 | 65.4 | 58.9 | 49.7 | 33.3 | 9.1 | 39.9 | 66.2 | 56.1 | 11.4 |
| **Ours-Extension** | - | **71.6** | **65.8** | **56.9** | **39.9** | **12.5** | **45.1** | **70.0** | **61.2** | **15.8** |

† Denotes additional optical flow input. Compared with our conference version, our extension model achieves significant performance gains and good efficiency.

with pixel-level masks of actors and actions for training and evaluating segmentation performance. We follow [11] and use its splits of 3,017 training videos, 737 testing videos and 28 unlabeled videos.

*J-HMDB Sentences [11]:* This dataset is an extension from the J-HMDB dataset [71] which contains 21 action categories, 928 videos and corresponding 928 referring expressions. For each video, one natural language referring expression is annotated to describe the actions performed by the actors. All the actors in the J-HMDB dataset are humans which are labeled with 2D articulated human puppet masks for segmentation evaluation.

*Refer-YouTube-VOS [18]:* This dataset is built upon the common video object segmentation dataset YouTube-VOS [45] which contains 3,978 video sequences with densely sampled (every 5 frames in 30-fps) multi-instance mask annotations and their corresponding referring expressions. Following the official description of collectors [18], we adopt the split of 3,471 training videos, 202 validation videos, and 305 testing videos.

*Refer-DAVIS [19]:* This dataset is extended from another common video object segmentation dataset DAVIS-17 [44] containing 60 training videos and 30 validation videos, where multiple instances are annotated pixel-wisely in each video. Based on the contents of the first frame and the whole video sequence, Refer-DAVIS annotates each video with two types of referring expressions respectively.

We follow prior works [11], [12] to adopt Precision@$X$ (P@$X$), Overall IoU, and Mean IoU as evaluation metrics. Overall IoU is defined as the ratio of the accumulated intersection area over the accumulated union area between ground-truth masks and predictions on all the test samples. Mean IoU measures the IoU between ground-truth masks and predictions averaged over all the test samples. Precision@$X$ calculates the percentage of test samples whose IoU are higher than a predefined threshold $X$, where $X \in [0.5, 0.6, 0.7, 0.8, 0.9]$. The mean Average Precision (mAP) [72] is also computed over the threshold section of $[0.50 : 0.05 : 0.95]$. For Refer-YouTube-VOS and Refer-DAVIS, we follow [18] to use region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) as metrics.

### B. Implementation Details

For the spatial encoder and temporal encoder, we adopt 2D ResNet [17] pretrained on the ImageNet [73] dataset and I3D [16] pretrained on the Kinetics400 [16] dataset as the backbone networks respectively. The spatial encoder takes the annotated target frame as input while the temporal encoder takes the video clip of 8 frames as input where the target frame is in the middle. We utilize GRU [65] as the language encoder to extract linguistic features where the hidden dims are set as 300. The GloVe word embedding [66] pretrained on the Common Crawl with 840B tokens is used to embed input words. The maximum sequence length of the input referring expression is set as 20. The height and width of the input frames are resized to $320 \times 320$. In the decoder, the number of feature channels $C_{\mathrm{d}} = 256$, and the output size of adaptive max pooling in the LASP module is set as $8 \times 8$. Adam [74] is utilized as the optimizer to train our model in an end-to-end manner and the model is trained for 15 epochs in total. For A2D Sentences and Refer-YouTube-VOS datasets, the initial learning rate is set as $1e^{-4}$ and we reduce it by $2\times$ after 10, 12, and 14 epochs. Following [11], [12], [15], we use the best model pretrained on the A2D Sentences dataset to evaluate it on the whole J-HMDB Sentences dataset without finetuning. For Refer-DAVIS, we adopt the best model pretrained on the Refer-YouTube-VOS dataset and finetune it for 1 epoch with the learning rate of $1e^{-5}$. All the inference speeds are calculated on the same machine with a single NVIDIA Tesla V100 GPU. Our method is implemented with PyTorch and MindSpore.

### C. Comparison With State-of-the-art Methods

We conduct experiments on four RVS datasets to compare our method with previous state-of-the-art methods. As shown in Table I, our methods outperform previous ones on the A2D Sentences testing set, indicating the effectiveness of spatial-temporal collaboration and adaptive visual-linguistic interaction. It is also worth mentioning that our extended method achieves significant performance gains over the previous conference version, which shows the language update in CMAM+, semantic propagation via language sampling of LASP and advanced end-to-end training strategy are quite effective improvements. For the most rigorous metric P@0.9, our extension is also superior to the performances of our conference version and other methods, demonstrating that our extended method can not only accurately identify the correct object through cross-modal alignment, but also generate a finer mask to cover the object. Since Mean IoU treats objects of different scales equally while

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE WHOLE J-HMDB SENTENCES DATASET USING THE BEST MODEL TRAINED ON A2D SENTENCES WITHOUT FURTHER FINETUNING

| Method | Pub. | Overlap | | | | | mAP | IoU | |
|---|---|---|---|---|---|---|---|---|---|
| | | P@0.5 | P@0.6 | P@0.7 | P@0.8 | P@0.9 | 0.5:0.95 | Overall | Mean |
| Gavrilyuk *et al.* [11] | CVPR18 | 69.9 | 46.0 | 17.3 | 1.4 | 0.0 | 23.3 | 54.1 | 54.2 |
| Gavrilyuk *et al.* ‡ [11] | CVPR18 | 71.2 | 51.8 | 26.4 | 3.0 | 0.0 | 26.7 | 55.5 | 57.0 |
| ACGA [12] | ICCV19 | 75.6 | 56.4 | 28.7 | 3.4 | 0.0 | 28.9 | 57.6 | 58.4 |
| VT-Capsule [13] | CVPR20 | 67.7 | 51.3 | 28.3 | 5.1 | 0.0 | 26.1 | 53.5 | 55.0 |
| CMDy [14] | AAAI20 | 74.2 | 58.7 | 31.6 | 4.7 | 0.0 | 30.1 | 55.4 | 57.6 |
| PRPE [15] | IJCAI20 | 69.0 | 57.2 | 31.9 | 6.0 | 0.1 | 29.4 | - | - |
| CMSA [43] | TPAMI21 | 76.4 | 62.5 | 38.9 | 9.0 | **0.1** | - | 62.8 | 58.1 |
| CMPC-V [42] | TPAMI21 | 81.3 | 65.7 | 37.1 | 7.0 | 0.0 | 34.2 | 61.6 | 61.7 |
| Ours [1] | CVPR21 | 78.3 | 63.9 | 37.8 | 7.6 | 0.0 | 33.5 | 59.8 | 60.4 |
| **Ours-Extension** | - | **86.5** | **75.2** | **49.8** | **12.2** | 0.0 | **39.2** | **64.0** | **65.2** |

‡ Denotes more layers of I3D backbone are trained on A2D Sentences. Our method shows notable generalization ability.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE REFER-YOUTUBE-VOS VALIDATION SET

| Method | Pub. | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|---|
| URVOS † [18] | ECCV20 | 41.34 | - | - |
| URVOS [18] | ECCV20 | 45.27 | 49.19 | 47.23 |
| CMPC-V [42] | TPAMI21 | 45.64 | 49.32 | 47.48 |
| **Ours-Extension** | - | **48.15** | **50.45** | **49.30** |

† Denotes multiple iterations of the second stage inference step are removed.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE REFER-DAVIS VALIDATION SET

| Method | Pub. | Pretrained | $\mathcal{J}\&\mathcal{F}$ | |
|---|---|---|---|---|
| | | | 1st Frame | Full Video |
| Khoreva *et al.* [19] | ACCV18 | RefCOCO [75] | 39.30 | 37.10 |
| URVOS [18] | ECCV20 | RefCOCO [75] | 44.10 | - |
| URVOS [18] | ECCV20 | Refer-YVOS [18] | 51.63 | - |
| **Ours-Extension** | - | Refer-YVOS [18] | **53.29** | **54.45** |

The results of two settings are provided.

Overall IoU favors large objects, our improvements on IoU metrics also show that our method can well handle the scale variation of objects. Our extension model also obtains faster inference speed than our conference version and prior methods due to optimized network implementations and structures, showing its good efficiency.

Following prior works [12], [14], [15], we further verify the generalization ability of our method on the whole J-HMDB Sentences dataset. The best model pretrained on the A2D Sentences dataset is adopted to directly evaluate all the samples on the J-HMDB Sentences without finetuning. As shown in the Table II, our extension also significantly outperforms previous state-of-the-art methods as well as our conference version, indicating that our method can excavate richer multimodal information through the collaborative learning of spatial and temporal encoders, leading to stronger generalization ability. Note that all the methods including ours yield approximate zero performance on P@0.9, which is probably because models cannot predict particularly complete masks on unseen samples without training on J-HMDB Sentences.

We also conduct more performance comparisons on two newly proposed datasets Refer-YouTube-VOS and Refer-DAVIS, in which video frames are annotated more densely and the object categories are richer as well. Table III summarizes the results on Refer-YouTube-VOS. Comparing with CMPC-V [42] and URVOS [18], our method achieves 1.82% and 2.07% improvements on $\mathcal{J}\&\mathcal{F}$ metric respectively, showing our method can well segment objects which are referred in complex scenes and long videos. Moreover, we adopt the best model pretrained on the Refer-YouTube-VOS dataset to finetune it on the Refer-DAVIS dataset following [18]. Results in Table IV also demonstrate that our model outperforms previous approaches.

### D. Ablation Studies

To evaluate the different designs of our framework, we conduct ablation studies on the A2D Sentences dataset.

*Component Analysis:* Table V presents the ablation results of our proposed encoders and modules. The 1-st row denotes the baseline model with only the temporal encoder (I3D [16]), where multimodal interaction only occurs in the decoder by visual and linguistic feature concatenation and fusion. The 2-nd row integrates spatial encoder (2D ResNet [17]) with temporal encoder to form a simple spatial-temporal collaboration model and achieves 3.3% and 2.0% performance improvements on mAP and Mean IoU metrics respectively, which demonstrates that introducing spatial encoder can supplement the temporal encoder with precise appearance information of the referred object and facilitate mask prediction. When inserting our proposed CMAM module in each stage of the two encoders, the performance in the 3-rd row obtains notable gains on all metrics, which shows the effectiveness of adaptive visual feature modulation with recombined language features. Moreover, by replacing CMAM with our improved version of CMAM+ in the 4th row, we can still observe performance boosts on all metrics, showing the additional language update path through the encoders can yield more comprehensive multimodal interactions. As shown in the last row, our proposed LASP module can yield further performance gains based on the model which

TABLE V
VERIFYING THE EFFECTIVENESS OF EACH COMPONENT IN OUR LANGUAGE-AWARE SPATIAL-TEMPORAL COLLABORATION FRAMEWORK

| Temporal | Spatial | CMAM | CMAM+ | LASP | Overlap |  |  |  |  | mAP | IoU |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  | P@0.5 | P@0.6 | P@0.7 | P@0.8 | P@0.9 | 0.5:0.95 | Overall | Mean |
| ✓ |  |  |  |  | 62.4 | 55.6 | 44.3 | 29.1 | 6.9 | 36.1 | 63.7 | 53.7 |
| ✓ | ✓ |  |  |  | 63.6 | 57.5 | 48.5 | 35.0 | 10.8 | 39.4 | 65.5 | 55.7 |
| ✓ | ✓ | ✓ |  |  | 67.6 | 62.1 | 53.5 | 38.6 | 12.1 | 42.6 | 67.3 | 58.5 |
| ✓ | ✓ |  | ✓ |  | 69.9 | 64.5 | 55.6 | 38.7 | 12.4 | 43.9 | 68.8 | 59.8 |
| ✓ | ✓ |  | ✓ | ✓ | **71.6** | **65.8** | **56.9** | **39.9** | **12.5** | **45.1** | **70.0** | **61.2** |

"Spatial" and "Temporal" denote the spatial encoder and temporal encoder respectively.

TABLE VI
INSERTING STAGES OF THE CMAM+ MODULE IN THE ENCODERS WITH LASP MODULE IN THE DECODER

| Stages |  |  |  |  | mAP | IoU |  | FPS |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 0.5:0.95 | Overall | Mean |  |
|  |  |  |  |  | 40.3 | 65.3 | 56.7 | 22.9 |
| ✓ |  |  |  |  | 40.4 | 65.6 | 57.0 | 21.2 |
| ✓ | ✓ |  |  |  | 42.7 | 67.6 | 58.8 | 19.8 |
| ✓ | ✓ | ✓ |  |  | 44.0 | 68.7 | 60.3 | 18.0 |
| ✓ | ✓ | ✓ | ✓ |  | 44.5 | 69.1 | 60.4 | 16.8 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **45.1** | **70.0** | **61.2** | 15.8 |

TABLE VII
INSERTING STAGES OF CMAM+ MODULE IN THE ENCODERS WITHOUT LASP MODULE IN THE DECODER

| Stages |  |  |  |  | mAP | IoU |  |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 0.5:0.95 | Overall | Mean |
|  |  |  |  |  | 39.4 | 65.5 | 55.7 |
| ✓ |  |  |  |  | 40.0 | 65.8 | 55.8 |
| ✓ | ✓ |  |  |  | 42.8 | 68.1 | 58.6 |
| ✓ | ✓ | ✓ |  |  | 43.6 | 68.2 | 59.5 |
| ✓ | ✓ | ✓ | ✓ |  | 43.7 | 68.5 | 59.7 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **43.9** | **68.8** | **59.8** |

TABLE VIII
INSERTING STAGES OF LASP MODULE

| Stages |  |  |  | mAP | IoU |  | FPS |
|---|---|---|---|---|---|---|---|
| 2→1 | 3→2 | 4→3 | 5→4 | 0.5:0.95 | Overall | Mean |  |
|  |  |  |  | 43.9 | 68.8 | 59.8 | 16.8 |
|  |  |  | ✓ | 44.4 | 69.2 | 60.6 | 16.6 |
|  |  | ✓ | ✓ | 44.5 | 69.4 | 61.0 | 16.2 |
|  | ✓ | ✓ | ✓ | **45.1** | **70.0** | **61.2** | 15.8 |
| ✓ | ✓ | ✓ | ✓ | 44.4 | 68.9 | 60.3 | 15.5 |

has already achieved high performance. This result indicates that propagating language-aware deep semantics to shallow stages can effectively identify the referred object and refine the segmentation mask as well.

*Inserting Positions of CMAM+:* We evaluate different inserting positions of the CMAM+ module with the LASP module in the decoder and present the results in Table VI. When CMAM+ is inserted into the 2-nd and 3-rd stages of spatial and temporal encoders, the model's performance can witness relatively significant improvements, which shows hierarchical cross-modal interactions beginning from the shallow stages can extract meaningful multimodal representations under the refinement of LASP. As we insert CMAM+ into the deeper stages of encoders, segmentation performance is also constantly improved, which shows visual features of different abstraction levels can be well modulated with dynamically recombined and progressively updated linguistic features. The influence on the inference speed in shallow stages is relatively larger than in deep stages since high-resolution features require more computations, but CMAM+ in shallow stages also obtains more notable performance gains. We also remove the LASP module to evaluate the independent performance gains of the CMAM+ module in different inserting positions. As shown in Table VIII, our CMAM+ modules also consistently yield performance gains in each stage with the same trend as in Table VI, and the accumulative improvements in all 5 stages are significant as well. Our CMAM+ modules in all 5 stages should be regarded as a whole, and the results in the two tables well demonstrate that our CMAM module can independently improve performance or collaborate with the LASP module for further improvement.

*Inserting Positions of LASP:* We also evaluate different inserting positions of the LASP module and summarize the results in Table VIII. Compared with the 1-st row, inserting LASP between the 5th and 4th stages can yield consistent performance improvements on mAP and IoU metrics. The same trend can be seen when inserting LASP between the 4th and 3-rd stages, which indicates the effectiveness of propagating deep language-aware semantics to shallow stages for highlighting foregrounds and suppressing backgrounds. In the 4th row of Table VIII, stacking LASP modules for three adjacent stages can further achieve 45.1% mAP and maintain high IoU performance, which also shows that aggregating and propagating features from the last three stages can obtain the balance between high-level semantics and low-level details, thus facilitating the model to identify more referred objects. However, the performance drops when incorporating features of the 1-st stage, which is probably because the semantic information of the shallowest stage is too inadequate and may introduce redundant noises. Our LASP module yields consistently slight effects on the inference speed in each stage of the decoder, which is because only a small set of pixels are sampled to propagate semantics between adjacent stages, and the computations are hence reduced.

*Sub-component Analysis of LASP:* We further verify the effectiveness of two sub-components of the LASP module, namely foreground feature highlighting and background feature suppression. The experimental results are summarized in Table IX. We can find that if our LASP module conducts only foreground feature highlighting by sampling and propagating semantics of language-compatible pixels from deep stages to shallow stages, the performance is improved accordingly. When the background feature suppression is introduced, it can still yield performance

TABLE IX
SUB-COMPONENT ANALYSIS OF LASP MODULE

| Method | mAP<br>0.5:0.95 | IoU<br>Overall | <br>Mean |
|---|---|---|---|
| w/o LASP | 43.9 | 68.8 | 59.8 |
| +FG only | 44.5 | 69.0 | 60.8 |
| +FG and BG (LASP) | **45.1** | **70.0** | **61.2** |

"FG" and "BG" denote foreground highlighting and background suppression, respectively.

TABLE X
COMPARATIVE EXPERIMENTS OF DIFFERENT LANGUAGE ENCODING METHODS ON A2D SENTENCES TESTING SET

| Dataset | Language Encoder | mAP<br>0.5:0.95 | IoU<br>Overall | <br>Mean |
|---|---|---|---|---|
| A2D Sentences | GRU [65] | **45.1** | **70.0** | **61.2** |
| | BERT [76] | 43.9 | 69.0 | 60.0 |
| | RoBERTa [77] | 42.6 | 67.7 | 58.0 |
| | XLNet [78] | 44.0 | 68.3 | 59.6 |

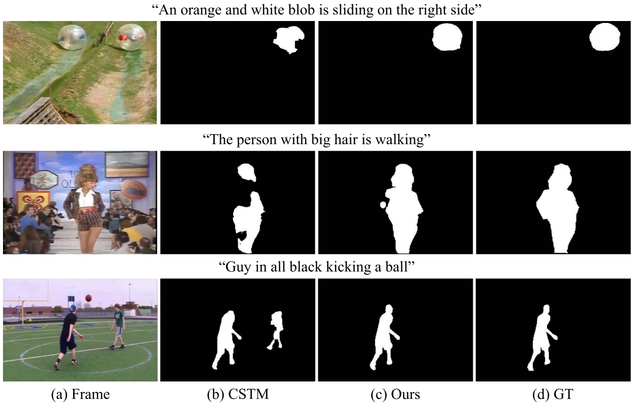

"An orange and white blob is sliding on the right side"
"The person with big hair is walking"
"Guy in all black kicking a ball"
(a) Frame
(b) CSTM
(c) Ours
(d) GT

Fig. 5. Qualitative comparison results. (a) Target frames. (b) Results of our previous conference version CSTM [1]. (c) Results of our extended method. (d) Ground-truth masks.

gains, which also indicates the effectiveness of sampling and propagating semantics of language-incompatible background pixels for noise reduction.

*Different Language Encoding Methods:* We conduct comparative experiments of our model using different language encoding methods on the A2D Sentences dataset. As shown in Table X, We can observe some counter-intuitive results where pretrained language models yield inferior performance than sequential model GRU. We suppose this phenomenon is probably because BERT-based large models may be more suitable for processing longer sentences or documents while GRU may work better at short sentences or phrases. Since the average length of referring expressions in A2D Sentences is 6.9 words, BERT may be too "heavy" to well handle this relatively shorter and less complicated corpus. In addition, Ezen-Can [79] conducts an empirical study that finds BERT does not always perform better than sequential models on different scenarios and corpus, which can also support our experimental results from the side.

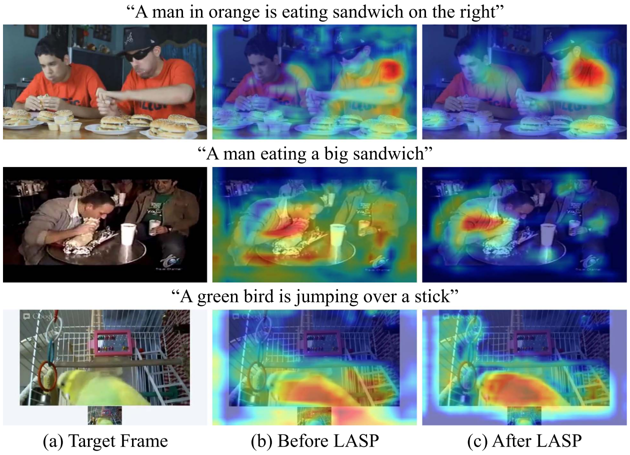

"A man in orange is eating sandwich on the right"
"A man eating a big sandwich"
"A green bird is jumping over a stick"
(a) Target Frame
(b) Before LASP
(c) After LASP

Fig. 6. Visualization of feature maps in LASP. (a) Target frames. (b) Feature activations before LASP modules. (c) Feature activations after LASP modules.

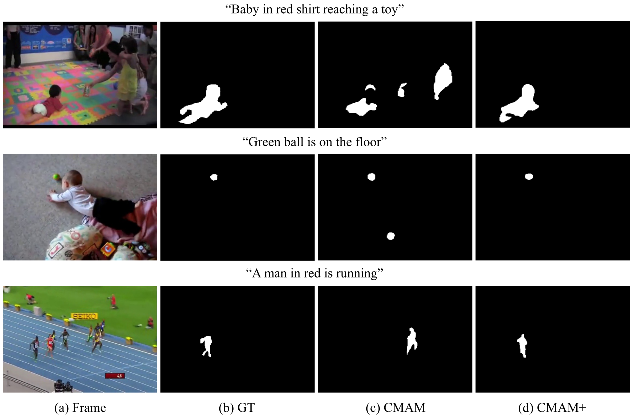

"Baby in red shirt reaching a toy"
"Green ball is on the floor"
"A man in red is running"
(a) Frame
(b) GT
(c) CMAM
(d) CMAM+

Fig. 7. Qualitative comparison between our CMAM and CMAM+ modules. (a) Target frames. (b) Ground-truth masks. (c) Results of our CMAM module (d) Results of our CMAM+ module.

## E. Qualitative Analysis

As illustrated in Fig. 5, we show qualitative comparison between our previous conference version CSTM [1] and our extended method. We can observe that our extended method can better recognize which object is performing the described action. Take the 3-rd row as an example, CSTM is confused about the two guys while our extended method can correctly identify the left guy using both appearance and motion clues and yield accurate segmentation, indicating the effectiveness of our extensions. In the 1-st and 2-nd rows, our extended method can also predict accurate segmentations on the referred objects under the existence of other distractors of the same categories (blobs and humans), showing the model's capacity in complex scenes.

We also visualize the feature maps of LASP modules in Fig. 6. Columns (b) and (c) show the feature activations before and after LASP modules respectively. We can observe that the application of LASP is able to highlight foreground feature regions that are compatible with the language description while suppressing background feature regions that are incompatible with the language description. For example, activations of the man eating

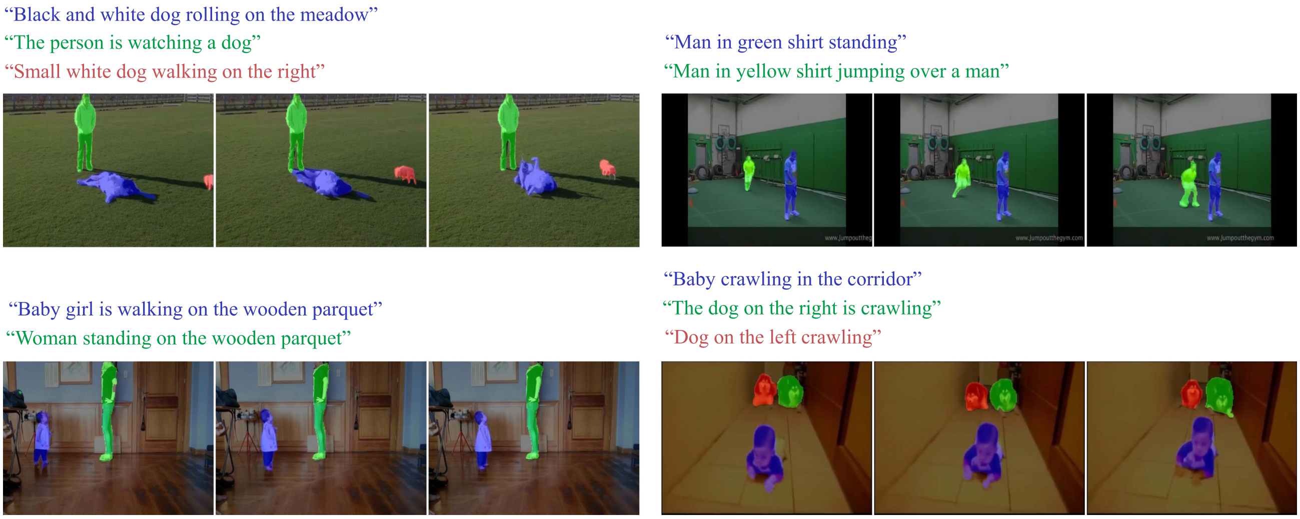

"Black and white dog rolling on the meadow"
"The person is watching a dog"
"Small white dog walking on the right"
"Man in green shirt standing"
"Man in yellow shirt jumping over a man"
"Baby girl is walking on the wooden parquet"
"Woman standing on the wooden parquet"
"Baby crawling in the corridor"
"The dog on the right is crawling"
"Dog on the left crawling"

Fig. 8. Qualitative results of consecutive frames on A2D Sentences testing set. The colors of referring expressions correspond to the colors of segmentation masks.

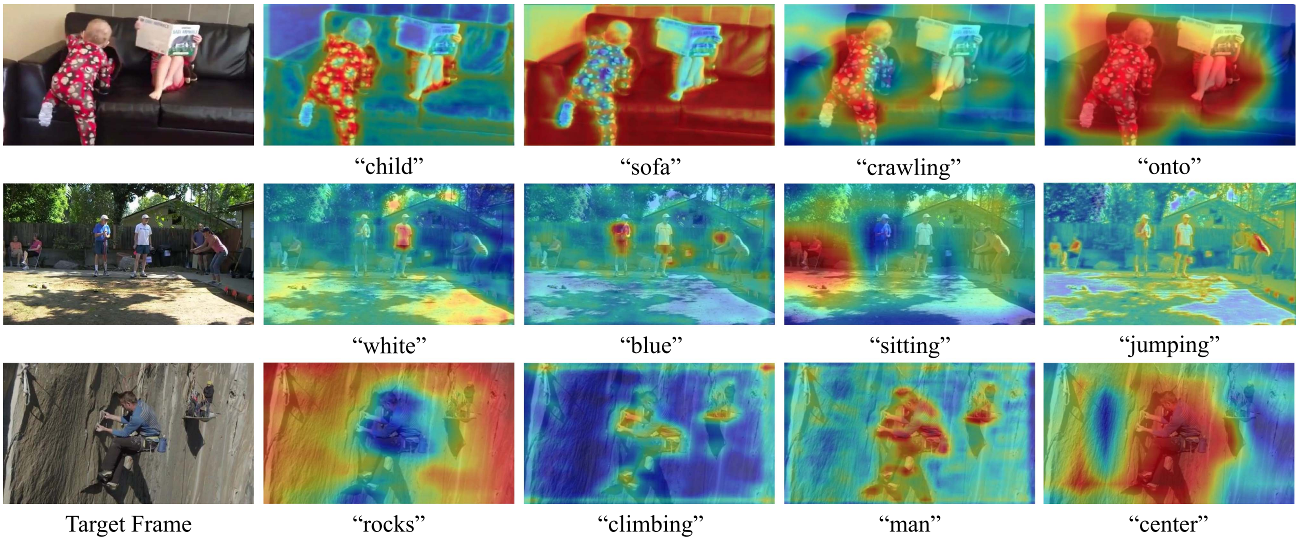


Fig. 9. Visualization of attention maps between words and frames in CMAM+. Red regions denote high attention values. Both spatial-related words and temporal-related words can attend to corresponding visual regions.

a big sandwich in the 2-nd row are highlighted while other background regions (e.g., the right sitting man and the table) are properly suppressed. Similar highlighting and suppression phenomenons can also be seen in the other two examples, which well demonstrate the effectiveness of LASP modules.

Fig. 7 presents the qualitative comparison results between our CMAM and CMAM+ modules. We can observe that our CMAM+ module can produce more accurate segmentation results than the CMAM module. For example, CMAM generates additional false predictions on other distractors with similar appearances such as the child in yellow in the 1-st row and the ball-like gadget in the 2-nd row. In the 3-rd row, CMAM misidentifies the referred man completely among these athletes but CMAM+ complements the man's head even though its annotation in the ground-truth image is missing. Our CMAM+ module can correctly identify and segment these referred objects in Fig. 7 with finer masks, showing the qualitative improvement obtained by CMAM+ over CMAM.

In Fig. 8, we show qualitative results of our extended method on consecutive multiple frames of the A2D Sentences testing set. Different colors of segmentation masks correspond to different colors of referring expressions. From these results, we can observe that our extended method is able to accurately segment different instances of the same categories (e.g., dogs and humans in these examples). With object movements on consecutive frames, our model can also predict stable segmentation results on the referred objects, which demonstrates the effectiveness of language-aware spatial-temporal collaboration.

In addition, we also visualize the cross-modal attention maps between visual features and linguistic features in our CMAM+ module. We show attention maps of different referring words in Fig. 9. Redder regions represent higher attention values. We can observe that both spatial-related words and temporal-related words can attend to corresponding visual regions with relatively higher attention values, which shows visual and linguistic features can be well associated in our CMAM+ module to better modulate and update features of different modalities. For example, the spatial-related words ("child", "white", "rocks") can correctly attend to the corresponding two children, the man in white clothes and the rocks on the cliffs. The temporal-related words ("crawling", "jumping", "climbing") can also attend to the corresponding objects performing these actions.

## V. Conclusion and Discussion

In this paper, we focus on the referring video segmentation (RVS) task which predicts the pixel-level mask of the object in the video referred by a natural language expression. Three main scientific problems are revealed for the RVS task, including spatial-temporal information exploitation, visual-linguistic cross-modal interaction, and multi-scale visual feature aggregation. First, successful RVS models are expected to sufficiently exploit both spatial and temporal information for correct distinguishment of the referred target in a video. However, previous methods only adopt 3D CNNs to extract an entangled spatio-temporal feature where misaligned spatial information from adjacent frames is introduced. To address this problem, we propose a language-aware spatial-temporal collaboration framework that provides undisturbed and precise spatial features of the referred target and recognizes the described actions respectively. Second, since both vision and language modalities are involved, RVS models are required to match the correct visual entity in the video with the linguistic semantic of referring expression. To this end, the CMAM and CMAM+ modules are proposed in the encoders to adaptively modulate visual features with recombined language features meanwhile updating language features with global contexts. Last, RVS models need to accurately aggregate multi-scale visual features for better segmenting objects with different sizes. Targeting at this issue, we propose an LASP module to conduct language-aware semantic propagation among multi-scale visual features, where language-compatible foreground pixels are highlighted and language-incompatible background pixels are suppressed. Extensive experiments on four RVS benchmarks show our method outperforms previous state-of-the-art methods.

*Limitation and Future Work:* In order to model temporal motion information, 3D CNN inevitably brings misaligned spatial information from adjacent frames. To alleviate this problem, we introduce an additional 2D CNN to compensate the target frame with undisturbed spatial information. In the future, we plan to explore new temporal motion modeling methods to reduce misaligned spatial information as much as possible. In addition, the cross-attention mechanism is a widely-adopted common practice to extract correlations between different modalities. In the future, we also plan to extract the visual-linguistic correlations with new techniques more tailored to the RVS task.

## Acknowledgements

A preliminary version of this work has appeared in CVPR 2021 [1].

## References

[1] T. Hui et al., "Collaborative spatial-temporal modeling for language-queried video actor segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4187–4196.

[2] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6819–6828.

[3] S. Jain, X. Wang, and J. E. Gonzalez, "Accel: A corrective fusion network for efficient semantic segmentation on video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8866–8875.

[4] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. Van Gool, "A survey on deep learning technique for video segmentation," 2021, *arXiv:2107.01153*.

[5] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 221–230.

[6] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9226–9235.

[7] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. V. Gool, "Video object segmentation with episodic graph memory networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 661–679.

[8] Y. Qi et al., "Reverie: Remote embodied visual referring expression in real indoor environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9982–9991.

[9] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7880–7889.

[10] G. Sreenu and M. Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, pp. 1–27, 2019.

[11] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek, "Actor and action video segmentation from a sentence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5958–5966.

[12] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3939–3948.

[13] B. McIntosh, K. Duarte, Y. S. Rawat, and M. Shah, "Visual-textual capsule routing for text-based video segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9942–9951.

[14] H. Wang, C. Deng, F. Ma, and Y. Yang, "Context modulated dynamic networks for actor and action video segmentation with language queries," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12152–12159.

[15] K. Ning, L. Xie, F. Wu, and Q. Tian, "Polar relative positional encoding for video-language segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, Art. no. 10.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] S. Seo, J.-Y. Lee, and B. Han, "URVOS: Unified referring video object segmentation network with a large-scale benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–223.

[19] A. Khoreva, A. Rohrbach, and B. Schiele, "Video object segmentation with language referring expressions," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 123–141.

[20] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 108–124.

[21] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1271–1280.

[22] R. Li et al., "Referring image segmentation via recurrent refinement networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5745–5753.

[23] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 630–645.

[24] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10502–10511.

[25] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7454–7463.

[26] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4424–4433.

[27] S. Huang et al., "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10488–10497.

[28] T. Hui et al., "Linguistic structure guided context modeling for referring image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 59–75.

[29] G. Luo et al., "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10034–10043).

[30] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, "Locate then segment: A strong pipeline for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9858–9867.

[31] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu, "Bottom-up shift and reasoning for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11266–11275.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017.

[33] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16321–16330.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[35] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "LAVT: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18155–18165.

[36] Z. Wang et al., "CRIS: Clip-driven referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11686–11695.

[37] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[38] Y. Rao et al., "DenseCLIP: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18082–18091.

[39] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso, "Can humans fly? action understanding with multiple classes of actors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2264–2273.

[40] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.

[41] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017.

[42] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4761–4775, Sep. 2022.

[43] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3719–3732, Jul. 2022.

[44] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.

[45] N. Xu et al., "Youtube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 585–601.

[46] Z. Yang, Y. Tang, L. Bertinetto, H. Zhao, and P. H. Torr, "Hierarchical interaction network for video object segmentation from referring expressions," in *Proc. Brit. Mach. Vis. Conf.*, 2021.

[47] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo, "Language as queries for referring video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4974–4984.

[48] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8741–8750.

[49] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4985–4995.

[50] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L.-C. Chen, "Feelvos: Fast end-to-end embedding learning for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9481–9490.

[51] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "RANet: Ranking attention network for fast video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3978–3987.

[52] H. Seong, S. W. Oh, J.-Y. Lee, S. Lee, S. Lee, and E. Kim, "Hierarchical memory matching network for video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12889–12898.

[53] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 2491–2502.

[54] H. Seong, J. Hyun, and E. Kim, "Kernelized memory network for video object segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 629–645.

[55] Z. Yang, Y. Wei, and Y. Yang, "Collaborative video object segmentation by multi-scale foreground-background integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4701–4712, Sep. 2022.

[56] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014.

[57] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[58] L. Wang et al., "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[59] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[60] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[61] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5533–5541.

[62] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[63] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7083–7093.

[64] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 909–918.

[65] K. Cho et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[66] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[67] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[69] S. Huang, Z. Lu, R. Cheng, and C. He, "FAPN: Feature-aligned pyramid network for dense image prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 864–873.

[70] X. Li et al., "PointFlow: Flowing semantics through points for aerial image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4217–4226.

[71] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3192–3199.

[72] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

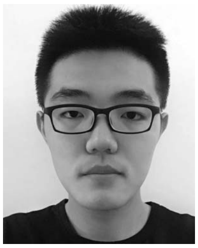
[75] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 792–807.

[76] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[77] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv: 1907.11692*.

[78] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[79] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus,", 2020, *arXiv: 2009.05451*.



**Tianrui Hui** received the BEng degree from Sun Yat-sen University. He is currently working toward the PhD degree with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include referring image/video segmentation and phrase grounding.

**Si Liu** received the PhD degree from the Institute of Automation, Chinese Academy of Sciences. She is currently a full professor with Beihang University. She has been a research assistant and postdoc in National University of Singapore. Her research interests include computer vision and multimedia analysis. She has published more than 40 cutting-edge papers on vision-language understanding, image/video segmentation and human parsing, etc. She was the recipient of the National Science Fund for Excellent Young Scholars. She has won the Best Paper Awards of ACM MM 2021 and 2013, the Best Demo Award of ACM MM 2012. She was the Champion of CVPR 2017 Look Into Person Challenge and the organizer of ECCV 2018, ICCV 2019 and CVPR 2021 Person in Context Challenges.

**Zihan Ding** received the BEng degree from Beihang University. He is currently working toward the PhD degree with the Institute of Artificial Intelligence, Beihang University. His research interests include referring expression segmentation and visual object tracking.

**Shaofei Huang** received the BS degree from Peking University. She is currently working toward the PhD degree with the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include scene parsing and visual grounding.

**Guanbin Li** received the BEng degree and the MS degree in computer science and technology from Sun Yat-sen University, in 2009 and 2012 respectively, and the PhD degree in computer science from the University of Hong Kong, advised by Prof. Yizhou Yu. He is currently an Associate Professor with Sun Yat-sen University. He has more than 70 publications in international conferences and journals, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Image Processing*, CVPR, ICCV, ICML, AAAI, and IJCAI, etc. His current research interests include computer vision, machine learning, and medical image analysis.

**Wenguan Wang** received the PhD degree from the Beijing Institute of Technology, in 2018. He is currently a lecturer and DECRA Fellow with University of Technology Sydney. From 2016 to 2018, he was a visiting PhD degree in University of California, Los Angeles. From 2018 to 2019, he was a Senior Scientist with the Inception Institute of Artificial Intelligence, UAE. From 2020 to 2022, he was a Research Fellow with ETH Zurich, Switzerland. He has published about 60 top journal and conference papers such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IJCV, NeurIPS, CVPR, ICCV, ECCV, and SIGGRAPH Asia. His current research interests include Neuro-Symbolic AI, Human-Centric AI, and Embodied AI. He is an associate editor of IEEE TCSVT and Neurocomputing.

**Luoqi Liu** received the PhD from the National University of Singapore (NUS), advised by Associate Professor Shuicheng Yan. He is currently Vice President of Meitu Inc. and Director of Meitu Image & Vision Lab. He is mainly working on multimedia, graphics and computer vision including video understanding, image editing and augmented reality. Till now, he has published more than 30 papers in top international journals and conferences, such as CVPR, NeurIPS, ECCV, ICCV, and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He also received ACM Multimedia Best Paper Award 2013 and PREMIA Best Student Paper Gold Prize 2014.

**Jizhong Han** received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences. He is currently a full professor with Institute of Information Engineering, Chinese Academy of Sciences. His research interests include multimedia information processing and big data storage. He has published more than 60 papers and held more than 10 domestic patents. He is also the principal investigator or participant of several National 973 or 863 Programs.