Spectrum-guided Multi-granularity Referring Video Object Segmentation

Bo Miao[1], Mohammed Bennamoun[1], Yongsheng Gao[2], Ajmal Mian[1]
[1]The University of Western Australia [2]Griffith University

Abstract

Current referring video object segmentation (R-VOS) techniques extract conditional kernels from encoded (low-resolution) vision-language features to segment the decoded high-resolution features. We discovered that this causes significant feature drift, which the segmentation kernels struggle to perceive during the forward computation. This negatively affects the ability of segmentation kernels. To address the drift problem, we propose a Spectrum-guided Multi-granularity (SgMg) approach, which performs direct segmentation on the encoded features and employs visual details to further optimize the masks. In addition, we propose Spectrum-guided Cross-modal Fusion (SCF) to perform intra-frame global interactions in the spectral domain for effective multimodal representation. Finally, we extend SgMg to perform multi-object R-VOS, a new paradigm that enables simultaneous segmentation of multiple referred objects in a video. This not only makes R-VOS faster, but also more practical. Extensive experiments show that SgMg achieves state-of-the-art performance on four video benchmark datasets, outperforming the nearest competitor by 2.8% points on Ref-YouTube-VOS. Our extended SgMg enables multi-object R-VOS, runs about 3× faster while maintaining satisfactory performance. Code is available at https://github.com/bo-miao/SgMg.

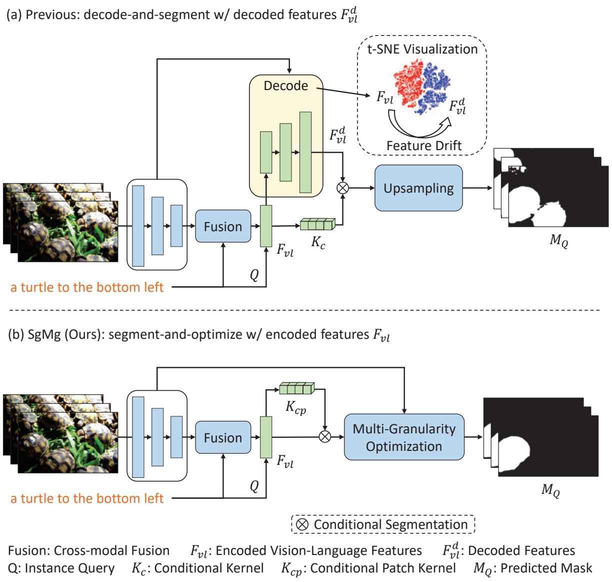

Figure 1. (a) Previous methods [4, 57] apply segmentation kernels $\mathcal{K}_c$ [50], extracted from encoded features $\mathcal{F}_{vl}$, to segment the decoded high-resolution features $\mathcal{F}_{vl}^d$. (b) We use segmentation kernels $\mathcal{K}_{cp}$, extracted from encoded features $\mathcal{F}_{vl}$, to segment the encoded features $\mathcal{F}_{vl}$ directly, and propose multi-granularity optimization to recover visual details and produce fine-grained masks.

1. Introduction

Referring video object segmentation (R-VOS) aims at segmenting objects in a video, referred to by linguistic descriptions. R-VOS is an emerging task for multimodal reasoning and promotes a wide range of applications, including language-guided video editing and human-machine interaction. Different from conventional semi-supervised video object segmentation [44, 7, 41], where the mask annotation for the first frame is provided for reference, R-VOS is more challenging due to the need for cross-modal understanding between vision and free-form language expressions.

Early R-VOS techniques [2, 21, 61] perform feature encoding, cross-modal interaction, and language grounding using convolutional neural networks (CNNs). However, the limited ability of CNNs to capture long-range dependencies and handle free-form features constrains the model performance. With the advancement of attention mechanisms [52, 43, 15, 16, 58], recent methods achieved significant improvement on R-VOS using cross-attention [49, 23, 27] for multimodal understanding and transformers [6, 56] for spatio-temporal representation. Based on transformers, conditional kernel [50] is then introduced to separate foreground from semantic features given its high adaptability to different instances [4, 57]. As illustrated in Fig. 1(a), these methods attend to encoded vision-language features $\mathcal{F}_{vl}$ using instance queries Q to predict conditional kernels $\mathcal{K}_c$, and employ $\mathcal{K}_c$ as the segmentation head to segment decoded features $\mathcal{F}_{vl}^d$. Despite the promising performance, this paradigm still has some limitations. *Firstly*, as shown in the t-SNE [51] visualization in Fig. 1(a), although the nonlinear decoding process introduces visual details, this is accompanied by a significant feature drift, which increases the difficulty of segmentation since $\mathcal{K}_c$ is predicted before feature decoding. *Secondly*, bilinear upsampling of the predicted masks $\mathcal{M}_Q$ to increase resolution impedes the segmentation performance. *Thirdly*, these methods only support single expression-based segmentation, making R-VOS inefficient when multiple referred objects exist in a video.

In this work, we propose a Spectrum-guided Multi-granularity (SgMg) approach that follows a segment-and-optimize pipeline to address the above problems. As depicted in Fig. 1(b), SgMg introduces Conditional Patch Kernel (CPK) $\mathcal{K}_{cp}$ to directly segment its fully perceived encoded features $\mathcal{F}_{vl}$, avoiding the feature drift and its adverse effects. The segmentation is then refined using our proposed Multi-granularity Segmentation Optimizer (MSO), which employs low-level visual details to produce full-resolution masks. Within the SgMg framework, we further develop Spectrum-guided Cross-modal Fusion (SCF) that performs intra-frame global interactions in the spectral domain to facilitate multimodal understanding. Finally, we introduce a new paradigm called multi-object R-VOS to simultaneously segment multiple referred objects in a video. To achieve this, we extend SgMg by devising multi-instance fusion and decoupling. Our main contributions are summarized as follows:

- We explain how existing R-VOS methods suffer from the feature drift problem. To address this problem, we propose SgMg that follows a *segment-and-optimize* pipeline and achieves top-ranked overall performance on multiple benchmark datasets.
- We propose Spectrum-guided Cross-modal Fusion to encourage intra-frame global interactions in the spectral domain.
- We extend SgMg to perform multi-object R-VOS, a new paradigm that enables simultaneous segmentation of multiple referred objects in a video. Our multi-object variant is more practical and runs $3\times$ faster.

We conduct extensive experiments on multiple benchmark datasets, including Ref-YouTube-VOS [49], Ref-DAVIS17 [21], A2D-Sentences [17], and JHMDB-Sentences [20], and achieve state-of-the-art performance on all four. On the largest validation set Ref-YouTube-VOS, SgMg achieves 65.7 $\mathcal{J}\&\mathcal{F}$ which is 2.8% points higher than that of the closest competitor ReferFormer [57]. On the A2D-Sentences, SgMg achieves 58.5 mAP which is 3.5% points higher than that of ReferFormer.

2. Related Works

Video Object Segmentation techniques fall into two categories: unsupervised and semi-supervised. Unsupervised approaches segment the most salient instances in each video without user interactions [38, 47]. They often employ two-stream networks to fuse motion and appearance cues for segmentation. Semi-supervised approaches track the given first frame object mask by performing online learning [5] or spatial-temporal association [44, 7, 66, 42, 53]. Unlike conventional semi-supervised video object segmentation, R-VOS takes a free-form linguistic expression as guidance to detect and segment referred objects in videos.

Referring Video Object Segmentation. R-VOS methods mainly use deep neural networks with vision-and-language interaction to empower visual features with corresponding linguistic information for pixel-level segmentation. For example, [49] employs a unified R-VOS framework that performs iterative segmentation guided by both language and temporal features. [32, 29] adopt progressive segmentation by perceiving potential objects and discriminating the best match. [70] fuses visual and motion features for segmentation under the guidance of linguistic cues. [30] models object relations to form tracklets and performs tracklet-language grounding. To enhance multi-modal interactions, [61, 14, 11, 10] perform hierarchical vision-language fusion on multiple feature layers.

Despite their promising performance, the complex multi-stage pipelines and use of multiple networks make R-VOS burdensome. To address these problems, MTTR [4] proposes an end-to-end transformer-based network with conditional kernels [50] to segment target objects. ReferFormer [57] further introduces language-guided instance queries to predict instance-aware conditional kernels and an auxiliary detection task to aid localization. These methods follow a decode-and-segment pipeline, which adopts conditional kernels to segment decoded high-resolution features to achieve promising performance. However, the nonlinear decoding process leads to significant feature drift that negatively affects the conditional kernels. In contrast to previous works, our approach follows a segment-and-optimize pipeline to avoid the adverse drift effects and to predict full-resolution masks in an efficient manner.

Vision and Language Representation Learning aims to learn vision-language semantics and alignment for multi-modal reasoning tasks. It has achieved significant success in various tasks [37, 67, 68], including video question answering [71], video captioning [1], video-text retrieval [12, 69], zero-shot classification [46], referring image/video segmentation [49], *etc*. Some approaches [46, 24] rely on contrastive pre-training using large-scale datasets to project different modalities into unified embedding space. Others [36, 13] develop cross-modal interaction layers for multimodal feature fusion and understanding. Recent deep

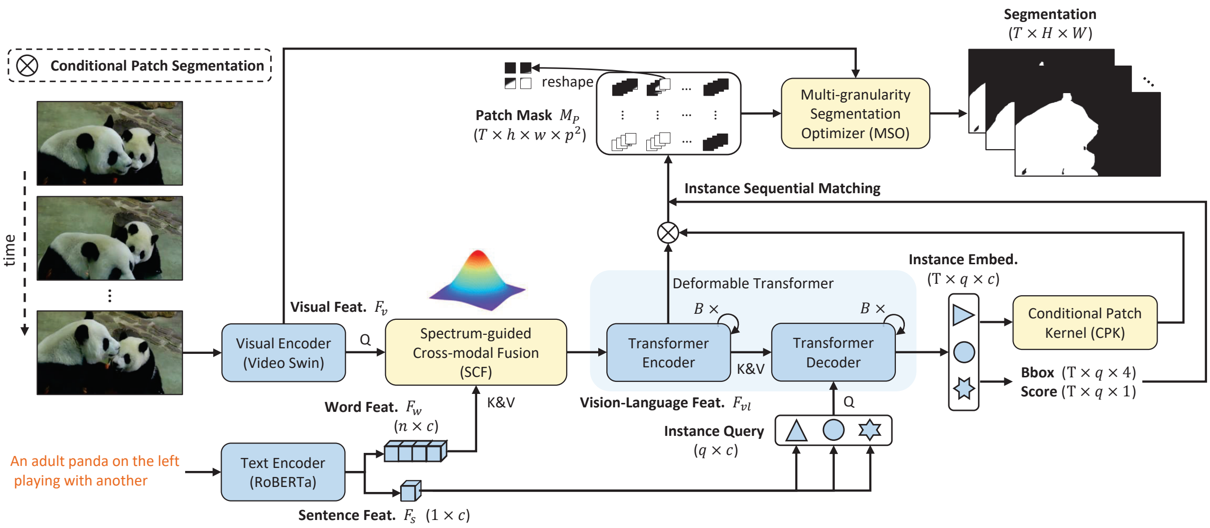

Figure 2. The overall framework of SgMg. Taking a video sequence $\mathcal{V} = \{I_i\}_{i=1}^{T}$ and a language expression $\mathcal{L} = \{S_i\}_{i=1}^{N}$ as input, SgMg predicts the masks of referred object $\mathcal{O}_{\mathcal{L}}$ in each frame. SCF projects visual features $\mathcal{F}_v$ to vision-language features $\mathcal{F}_{vl}$, instance-aware CPK predicts patch masks by segmenting encoded $\mathcal{F}_{vl}$, and MSO optimizes patch masks to get fine-grained results.

learning methods in spectral domain [18, 40, 8, 34] have raised widespread awareness because of their ability to perform global interactions. We take inspiration from these spectral-based methods and employ spectrum guidance in the field of vision-language representation to encourage multimodal global interactions.

3. SgMg: Spectrum-guided Multi-granularity Referring Video Object Segmentation

Given a video sequence $\mathcal{V} = \{I_i\}_{i=1}^{T}$ with T frames and a language query $\mathcal{L} = \{S_i\}_{i=1}^{N}$ with N words. The goal of R-VOS is to segment the referred object $\mathcal{O}_{\mathcal{L}}$ in $\mathcal{V}$ at pixel-level. To this end, we introduce a new approach termed SgMg. Different from previous R-VOS methods [4, 57], our approach follows a segment-and-optimize pipeline.

An overview of SgMg is shown in Fig. 2. Video Swin Transformer [35] is adopted to extract visual feature $\mathcal{F}_v$ and RoBERTa [33] is adopted to extract sentence $\mathcal{F}_s$ and word $\mathcal{F}_w$ features. The channel dimension of all features is projected to 256. Spectrum-guided Cross-modal Fusion (SCF) cross attends $\mathcal{F}_v$ with $\mathcal{F}_w$ to compute vision-language features $\mathcal{F}_{vl}$. Deformable Transformer [72] encoder is used to encode $\mathcal{F}_{vl}$ and the decoder associates instance queries created based on $\mathcal{F}_s$ to predict instance embeddings and the corresponding Conditional Patch Kernels (CPKs). Finally, the CPKs are employed to segment $\mathcal{F}_{vl}$ and predict patch masks that are further optimized with visual details through Multi-granularity Segmentation Optimizer (MSO). The choice of the encoder and transformer follows previous works to avoid distractions [4, 57].

3.1. Feature Drift Analysis

Existing R-VOS methods [57, 4] follow a decode-and-segment pipeline where conditional kernels $\mathcal{K}_c$ [50] are extracted from encoded features $\mathcal{F}_{vl}$ and used to segment the decoded features $\mathcal{F}_{vl}^{d}$. However, the decoding process leads to feature drift, which is evident in the t-SNE visualization depicted in Fig. 1(a). This drift is difficult for the kernels $\mathcal{K}_c$ to perceive during the forward computation since $\mathcal{K}_c$ is predicted before the feature decoding. Therefore, we argue that *even though the feature decoding enhances visual details, it also causes the drift problem that negatively affects the segmentation kernels*. This makes the existing decode-and-segment pipeline sub-optimal.

To overcome the adverse effects of feature drift while recovering visual details, we present SgMg, a novel approach that follows a *segment-and-optimize* pipeline. In a nutshell, SgMg performs Spectrum-guided Cross-modal Fusion to compute $\mathcal{F}_{vl}$, leverages Conditional Patch Kernels to segment encoded features $\mathcal{F}_{vl}$ to avoid the drift effects, and recovers visual details with Multi-granularity Segmentation Optimizer to generate fine-grained masks.

3.2. Spectrum-guided Cross-modal Fusion

The two-dimensional discrete Fourier transform converts spatial data into the spectral domain. Based on the spectral convolution theorem [3], point-wise update of signals in the spectral domain globally affects all inputs in the spatial domain, which gives the insight to design spectrum-based modules so as to efficiently facilitate global interactions, which is critical for multimodal understanding. In addition, Low-frequency components in the spectral domain usually

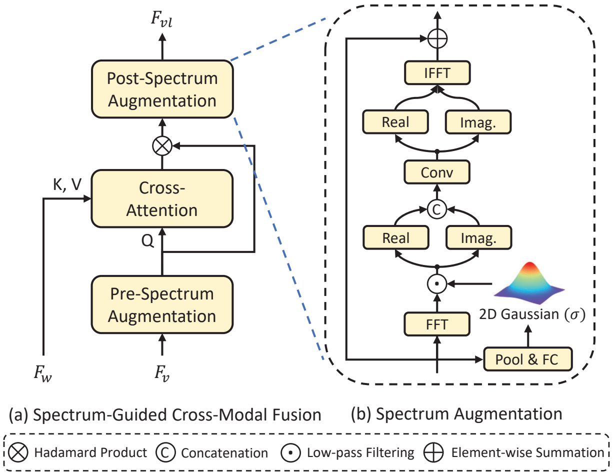

Figure 3. Spectrum-guided Cross-modal Fusion. Imag.: Imaginary. Pre-spectrum augmentation and post-spectrum augmentation share an identical structure.

correspond to the general semantic information according to previous theoretical studies [59, 60, 64].

Inspired by the above observations, we conjecture that low-frequency components can benefit higher dimensional semantic features and propose Spectrum-guided Cross-modal Fusion (SCF). As shown in Fig. 3, SCF performs pre-spectrum augmentation to enhance visual features before cross-modal fusion and post-spectrum augmentation to facilitate global vision-language interactions after the fusion process. Let $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$ denotes the input features, the spectrum augmentation (SA) is computed as:

$$\mathrm{SA}(\mathcal{F}, K) = \mathcal{F} + \Theta_{IFFT}(\mathrm{Conv}(\sigma(K, \mathcal{F}) \odot \Theta_{FFT}(\mathcal{F}))) \quad (1)$$

where $\odot$ denotes low-pass filtering with adaptive Gaussian smoothed filters $\sigma(K, \mathcal{F})$, which has the same spatial size as $\mathcal{F}$, and K is the bandwidth. To make $\sigma(K, \mathcal{F})$ input-aware, we create an initial 2D Gaussian map based on K, and apply pooling and linear layers on $\mathcal{F}$ to predict a scale parameter to update the Gaussian map. Thanks to the spectral convolution theorem, the efficient point-wise spectral convolution globally updates $\mathcal{F}$. We treat the spectral-operated features as residuals and add them to the original input features for enhancement. Overall, SCF, which takes visual features $\mathcal{F}_v$ and word-level text features $\mathcal{F}_w$ as input, is computed as:

$$\mathrm{SCF}(\mathcal{F}_w, \mathcal{F}_v) = \mathrm{SA}(\mathrm{SA}(\mathcal{F}_v) \otimes \mathrm{Att}(\mathrm{SA}(\mathcal{F}_v), \mathcal{F}_w)) \quad (2)$$

3.3. Conditional Patch Segmentation

We devise Conditional Patch Kernel (CPK) as the segmentation head to predict patch masks from the encoded vision-language features $\mathcal{F}_{vl}$ that are fully perceived by CPK. Unlike previous works [4, 57], CPK predicts a sequence of labels for each token rather than a single label, efficiently improving segmentation resolution along the channel dimension.

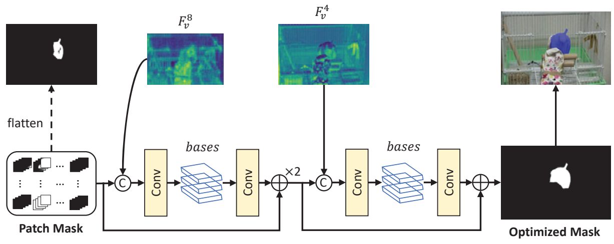

Figure 4. Multi-granularity Segmentation Optimizer, which predicts residual maps to optimize patch masks $\mathcal{M}_P$ progressively. Flatten: reshape $\mathcal{M}_P$ from $\mathbb{R}^{\frac{H}{i}\times\frac{W}{i}\times p^2}$ to $\mathbb{R}^{\frac{Hp}{i}\times\frac{Wp}{i}}$ for visualization. $\times 2$: upsampling operation.

Specifically, we first use sentence-level text features $\mathcal{F}_s$ and multiple learnable embeddings to generate instance queries $Q \in \mathbb{R}^{N\times C}$. Next, Q is projected into instance embeddings $\mathcal{E} \in \mathbb{R}^{N\times C}$ using the transformer decoder and $\mathcal{E}$ is leveraged to predict CPK for each instance query:

$$\mathrm{CPK}(Q, \mathcal{F}_{vl}) = \Theta(\mathrm{FC}(\mathrm{Att}(Q, \mathcal{F}_{vl}))) \quad (3)$$

where Θ denotes the parameterization operation that reshapes CPK to form two point-wise convolutions with the output channel number of 16, which is similar to [50]. Since Q changes dynamically according to different linguistic expressions, CPK becomes instance-aware and can separate objects of interest from F_{vl}. Finally, we apply the parameterized CPK (dynamic point-wise convolutions) on F_{vl} to predict patch masks $\mathcal{M}_P \in \mathbb{R}^{\frac{H}{i}\times\frac{W}{i}\times p^2}$, where $\frac{H}{i} \times \frac{W}{i}$ denotes the spatial resolution of F_{vl} and p^2 denotes the increased segmentation resolution on the channel dimension.

During inference, we can reshape patch masks to $\mathcal{M}_P \in \mathbb{R}^{\frac{Hp}{i}\times\frac{Wp}{i}}$ to efficiently generate fine-grained segmentation from low-resolution F_{vl}. The resolution of prediction will be consistent with the input when p equals to i. We found that this efficient CPK can achieve competitive performance compared to methods that use heavy decoders.

3.4. Multi-granularity Segmentation Optimizer

Segmenting encoded features $\mathcal{F}_{vl}$ with CPK avoids the detrimental drift effect on the segmentation head. However, visual details are required to produce accurate fine-grained masks. We propose Multi-granularity Segmentation Optimizer (MSO) to achieve this goal.

An overview of MSO is shown in Fig. 4. It takes the predicted patch masks $\mathcal{M}_P$ as object priors and reuses visual features $\mathcal{F}_v$ with spatial strides of {4,8} to gradually recover visual details and refine the priors. Specifically, MSO first concatenates $\mathcal{M}_P$ and $\mathcal{F}_v$ and projects them to low dimensional bases. Next, residual masks predicted by performing another convolution on these bases are used to correct $\mathcal{M}_P$. Finally, the optimized patch masks achieve the input resolution by reshaping from $\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4^2}$ to $\mathbb{R}^{H\times W}$. Since

Methods	Single-frame	Multi-frames	Multi-objects
[49, 63, 11] *et al.*	✓		
[4, 57, 56, 23] *et al.*	✓	✓	
Fast SgMg (Ours)	✓	✓	✓

Table 1. Comparing different methods for their ability to segment single or multiple frames or multiple objects simultaneously.

MSO does not include heavy computations, the segment-and-optimize pipeline makes our approach perform better with efficient inference time.

3.5. Multi-Object R-VOS

Existing R-VOS methods perform single-frame (frame-wise) segmentation [49, 63, 11] or multi-frame (clip-wise) segmentation [4, 57, 56] for an *individual* referred object at a time. However, to the best of our knowledge, no existing work explores the simultaneous segmentation of *multiple* referred objects in video using a common GPU, which is important for real-world scenarios. To fill this gap, we present a new paradigm called multi-object R-VOS.

The key to multi-object R-VOS is designing a network that shares computationally intensive features for multiple objects, and enables different instance features to be decoupled before segmentation. To achieve this, we extend SgMg for multi-object R-VOS by introducing multi-instance fusion and decoupling. As shown in Table 1, our method, dubbed Fast SgMg, can simultaneously segment multiple objects (in multiple frames) using a single 24GB GPU.

Fast SgMg shares visual features as well as vision-language features for all referred objects to make the network efficient, and decouples the shared features to make them instance-specific before the segmentation stage. Firstly, visual features ($\mathcal{F}_v$) and language features ($\mathcal{F}_w$ and $\mathcal{F}_s$) are extracted. Next, we associate $\mathcal{F}_v$ and $\mathcal{F}_w$ using multi-instance fusion rather than the previous SCF. Multi-instance fusion is built on the foundation of SCF, which is depicted in Fig. 3. The difference is that multi-instance fusion includes semantic fusion, which performs an element-wise add operation, after cross-attention to merge vision-language features of different expressions. The features after semantic fusion perform Hadamard product with $\mathcal{F}_v$ to generate the vision-language features $\mathcal{F}_{vl}$ for all objects:

$$\mathrm{SF}(\mathcal{F}_w, \mathcal{F}_v) = \sum_{i=1}^{N} \mathrm{Att}(\mathcal{F}_w^i, \mathcal{F}_v) \tag{4}$$

$$\mathrm{MIF}(\mathcal{F}_w, \mathcal{F}_v) = \mathrm{SA}(\mathrm{SA}(\mathcal{F}_v) \otimes \mathrm{SF}(\mathcal{F}_w, \mathrm{SA}(\mathcal{F}_v))) \tag{5}$$

where $\otimes$ denotes Hadamard product and N denotes the number of expressions. After vision-language fusion, we encode $\mathcal{F}_{vl}$ using the transformer encoder to enrich its semantic information, and plug multi-instance decoupling to decouple features for each instance. Multi-instance decoupling employs $\mathcal{F}_w$ and cross-attention to decouple $\mathcal{F}_{vl}$ to predict instance embeddings $\mathcal{E}$ for different referred objects. These embeddings are then projected to CPKs to predict the patch masks. Thus, FAST SgMg shares features, which account for most of the computational overhead, for different expressions, making it efficient for referring segmentation.

3.6. Instance Matching and Loss Functions

Following [4, 57], we perform instance matching with $N = 5$ learnable instance queries to improve fault tolerance. These queries are projected to CPKs to predict N potential patch masks $\mathcal{M}_P$ for each expression. The Hungarian algorithm [22] is then adopted to select the best match based on the matching loss for training. During inference, we directly employ the predicted confidence scores $\mathcal{S}$ to measure the instance queries and select the results.

We adopt the same training losses and weights as used in [57, 72] for a fair comparison. Specifically, we use Dice loss [26] and Focal loss [31] for patch mask $\mathcal{M}_P$ and optimized mask $\mathcal{M}_O$, Focal loss [31] for confidence scores $\mathcal{S}$, and L1 and GIoU [48] loss for bounding boxes $\mathcal{B}$. The final training loss functions are:

$$\mathcal{L}_{train} = \lambda_{\mathcal{M}_P}\mathcal{L}_{\mathcal{M}_P} + \lambda_{\mathcal{M}_O}\mathcal{L}_{\mathcal{M}_O} + \lambda_{\mathcal{B}}\mathcal{L}_{\mathcal{B}} + \lambda_{\mathcal{S}}\mathcal{L}_{\mathcal{S}} \tag{6}$$

where $\mathcal{L}$ and λ are the loss term and weight, respectively.

4. Experiments

4.1. Datasets and Metrics

Datasets. We evaluate SgMg on four video benchmarks: Ref-YouTube-VOS [49], Ref-DAVIS17 [21], A2D-Sentences [17], and JHMDB-Sentences [20]. Ref-YouTube-VOS is currently the largest dataset for R-VOS, containing 3,978 videos with about 13K expressions. Ref-DAVIS17 is an extension of DAVIS17 [45] by including the language expressions of different objects and contains 90 videos. A2D-Sentences is a general actor and action segmentation dataset with over 3.7K videos and 6.6K action descriptions. JHMDB-Sentences includes 928 videos and 928 descriptions covering 21 different action classes.

Evaluation Metrics. We adopt the standard metrics to evaluate our models: region similarity $\mathcal{J}$ (average IoU), contour accuracy $\mathcal{F}$ (average boundary similarity), and their mean value $\mathcal{J}\&\mathcal{F}$. All results are evaluated using the official code or server. On A2D-Sentences and JHMDB-Sentences, we adopt mAP, overall IoU, and mean IoU for evaluation.

4.2. Implementation Details

Following [4, 6, 57], we train our models on the training set of Ref-YouTube-VOS, and directly evaluate them on the validation split of Ref-YouTube-VOS and Ref-DAVIS17 without any additional techniques, *e.g.*, model ensemble, joint training, and mask propagation, since they are not the focus of this paper. Additionally, we present results for our

Method	Year	Backbone	Ref-YouTube-VOS				Ref-DAVIS17		
			$\mathcal{J}$&$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS	$\mathcal{J}$&$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [63]	2019	ResNet-50	36.4	34.8	38.1	-	40.2	36.9	43.5
URVOS [49]	2020	ResNet-50	47.2	45.3	49.2	-	51.5	47.3	56.0
CMPC-V [32]	2021	I3D	47.5	45.6	49.3	-	-	-	-
PMINet [11]	2021	ResNeSt-101	53.0	51.5	54.5	-	-	-	-
YOFO [23]	2022	ResNet-50	48.6	47.5	49.7	10	53.3	48.8	57.8
LBDT [10]	2022	ResNet-50	49.4	48.2	50.6	-	54.3	-	-
MLRL [56]	2022	ResNet-50	49.7	48.4	51.0	-	52.8	50.0	55.4
MTTR [4]	2022	Video-Swin-T	55.3	54.0	56.6	-	-	-	-
MANet [6]	2022	Video-Swin-T	55.6	54.8	56.5	-	-	-	-
ReferFormer [57]	2022	Video-Swin-T	56.0	54.8	57.3	50	-	-	-
SgMg (Ours)	2023	Video-Swin-T	**58.9**	**57.7**	**60.0**	**65**	**56.7**	**53.3**	**60.0**
Pre-training with RefCOCO/+/g & larger backbone									
ReferFormer [57]	2022	Video-Swin-T	59.4	58.0	60.9	50	59.6	56.5	62.7
SgMg (Ours)	2023	Video-Swin-T	**62.0**	**60.4**	**63.5**	**65**	**61.9**	**59.0**	**64.8**
ReferFormer [57]	2022	Video-Swin-B	62.9	61.3	64.6	33	61.1	58.1	64.1
SgMg (Ours)	2023	Video-Swin-B	**65.7**	**63.9**	**67.4**	**40**	**63.3**	**60.6**	**66.0**

Table 2. Quantitative comparison to state-of-the-art methods on the validation split of Ref-YouTube-VOS and Ref-DAVIS17.

Method	Backbone	A2D-Sentences			JHMDB-Sentences		
		mAP	Overall IoU	Mean IoU	mAP	Overall IoU	Mean IoU
Hu *et al.* [19]	VGG-16	13.2	47.4	35.0	17.8	54.6	52.8
Gavrilyuk *et al.* [17]	I3D	19.8	53.6	42.1	23.3	54.1	54.2
ACAN [54]	I3D	27.4	60.1	49.0	28.9	57.6	58.4
CMPC-V [32]	I3D	40.4	65.3	57.3	34.2	61.6	61.7
ClawCraneNet [29]	ResNet-50/101	-	63.1	59.9	-	64.4	65.6
MTTR [4]	Video-Swin-T	46.1	72.0	64.0	39.2	70.1	69.8
ReferFormer [57]	Video-Swin-T	52.8	77.6	69.6	42.2	71.9	71.0
SgMg (Ours)	Video-Swin-T	**56.1**	**78.0**	**70.4**	**44.4**	**72.8**	**71.7**
ReferFormer [57]	Video-Swin-B	55.0	78.6	70.3	43.7	73.0	71.8
SgMg (Ours)	Video-Swin-B	**58.5**	**79.9**	**72.0**	**45.0**	**73.7**	**72.5**

Table 3. Quantitative comparison to state-of-the-art R-VOS methods on A2D-Sentences and JHMDB-Sentences.

models first pre-trained on RefCOCO/+/g [39, 65] and then fine-tuned on Ref-YouTube-VOS. Similar to [57, 72], we set the coefficients for different losses λ_{dice}, λ_{focal}, λ_{L1}, λ_{giou} to 5, 2, 5, 2, respectively. The models are trained using 2 RTX 3090 GPUs with 5 frames per clip for 9 epochs. All frames are resized to have the longest side of 640 pixels. Further implementation details are in the supplementary material.

4.3. Quantitative Results

Ref-YouTube-VOS and Ref-DAVIS17. We compare SgMg with recently published works in Table 2. Our approach surpasses present solutions on the two datasets across all metrics. On Ref-YouTube-VOS, SgMg with the Video Swin Tiny backbone achieves 58.9 $\mathcal{J}$&$\mathcal{F}$ at 65 FPS, which is **2.9%** higher and **1.3**× faster than the previous state-of-the-art ReferFormer [57]. Our approach runs faster due to the use of the segment-and-optimize pipeline, which avoids the need for heavy feature decoders. When pre-training with RefCOCO/+/g and using a larger backbone, *i.e.*, Video Swin Base, the performance of SgMg further boosts to 65.7 $\mathcal{J}$&$\mathcal{F}$, consistently leading all other solutions by more than **2.8%**. On Ref-DAVIS17, SgMg achieves 63.3 $\mathcal{J}$&$\mathcal{F}$, outperforming state-of-the-art by **2.2%** and demonstrating the generality of our approach.

A2D-Sentences and JHMDB-Sentences. We further evaluate SgMg on A2D-Sentences and JHMDB-Sentences in Table 3. Following [57], the models are first pre-trained on RefCOCO/+/g and then fine-tuned on A2D-Sentences. JHMDB-Sentences is used only for evaluation. As shown in Table 3, SgMg achieves superior performance compared to other state-of-the-art R-VOS methods and surpasses the nearest competitor Referformer [57] by **3.5/1.3%** mAP on A2D-Sentences and JHMDB-Sentences, respectively.

Multi-object R-VOS. We extend SgMg to perform multi-object R-VOS, which is more practical and efficient for deployment. Fast SgMg is trained on Ref-YouTube-VOS without pre-training or postprocessing techniques. We

Method	Ref-DAVIS17			Ref-YouTube-VOS			
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
ReferFormer [57]	54.5	51.0	58.0	56.0	54.8	57.3	50
Fast SgMg (Ours)	54.2	51.1	57.3	54.2	53.1	55.3	**185**

Table 4. Evaluation of Fast SgMg on Ref-DAVIS17 and Ref-YouTube-VOS. Video-Swin-T is adopted as the backbone.

Components			Performance			
CPK	MSO	SCF	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
			54.4	52.7	56.2	70
✓			55.8	54.5	57.1	70
	✓		57.7	56.3	59.1	69
		✓	55.8	54.3	57.4	66
✓	✓		57.9	56.7	59.1	69
✓	✓	✓	58.9	57.7	60.0	65

Table 5. Ablation of different components on Ref-YouTube-VOS.

benchmark Fast SgMg on Ref-YouTube-VOS and Ref-DAVIS17 using the commonly used Video Swin Tiny, and compare the results with the state-of-the-art R-VOS method, which performs single-object segmentation.

As shown in Table 4, Fast SgMg achieves reasonable performance and runs about **3.7**× faster (**185** *vs* 50 FPS) compared to ReferFormer [57]. It should be noted that each object in the above datasets contains multiple expressions. On Ref-DAVIS17, we group expressions to have only one expression per object within each group and segment all expressions in each group simultaneously since the object identity is given. On Ref-YouTube-VOS, all expressions in a video are segmented simultaneously due to the lack of object identity, making it more challenging.

4.4. Ablation Study for Different Components

We conduct ablation experiments to evaluate the effectiveness of different components in SgMg. The components are added to the baseline model step-by-step.

Conditional Patch Kernel. As shown in Table 5, CPK boosts the performance by 1.4% compared with the recent instance-aware conditional kernels [57]. The sequential labels of each token predicted by CPK contain more fine-grained information, making the prediction more accurate.

Multi-granularity Segmentation Optimizer. We devise MSO to optimize the predicted patch masks. As shown in Table 5, MSO improves the performance by 3.3%, indicating the importance of fine-grained visual details in R-VOS.

Spectrum-guided Cross-modal Fusion. We present SCF to perform global interactions by operating in the spectral domain. In Table 5, using SCF to replace the traditional cross-attention in [57, 52] improves the $\mathcal{J}\&\mathcal{F}$ by 1.4%. We consider SCF extracts important low-frequency features and facilitates multimodal understanding globally, which is suitable for R-VOS since locating referred objects requires understanding the global context and token relations.

Settings	Drift	Pipeline	$\mathcal{J}\&\mathcal{F}$	FPS
Baseline + Decoder	✓	decode-and-segment	56.0	50
Baseline + Decoder + MSO	✓	decode-and-segment	56.4	49
Baseline + MSO	×	segment-and-optimize	**57.7**	**69**

Table 6. Feature drift analysis using ReferFormer [57] (Baseline + Decoder) and SgMg w/o CPK & SCF (Baseline + MSO). Significant improvement is achieved by addressing the drift issue (last row). Adding MSO on top of ReferFormer to recover visual details (for a second time) still performs worse than our basic pipeline.

Method	RefCOCO	RefCOCO+	RefCOCOg
MaIL [28]	70.1	62.2	62.5
CRIS [55]	70.5	62.3	59.9
RefTR [25]	70.6	-	-
LAVT [62]	72.7	62.1	61.2
VLT [9]	73.0	63.5	63.5
SgMg (Ours)	**76.3**	**66.4**	**70.0**

Table 7. Quantitative evaluation on the validation split of RefCOCO/+/g. Overall IoU is adopted as the evaluation metric.

4.5. Ablation Study for Feature Drift

We conduct ablation study in Table 6 to demonstrate the feature drift problem. Our segment-and-optimize pipeline addresses the adverse drift effect discussed in Section 3.1 to significantly outperform ReferFormer [57] by 1.7% points and runs 1.4× faster. Furthermore, adding MSO on top of ReferFormer still performs worse due to the negative drift impact caused by the decode-and-segment pipeline. These results demonstrate the efficacy of our proposed segment-and-optimize pipeline.

4.6. Referring Image Segmentation Results

We apply SgMg to referring image (expression) segmentation without any architectural modifications, and compare against the current state-of-the-art methods on RefCOCO/+/g [39, 65]. A single SgMg model is trained on RefCOCO/+/g without large-scale pre-training. As shown in Table 7, SgMg achieves advanced performance on all three benchmarks. These results demonstrate the efficacy of SgMg in referring image segmentation.

4.7. Inference Time Analysis of Multi-Object RVOS

We analyze the efficiency of the proposed multi-object R-VOS paradigm by comparing the FPS of Fast SgMg and SgMg on videos with different numbers of expressions. As illustrated in Fig. 6, Fast SgMg performs about 2× faster than SgMg when there are two expressions per video on average. As the number of expressions increases, Fast SgMg achieves faster reasoning time per object per frame due to its utilization of the multi-object R-VOS paradigm. When there are ten expressions in each video, Fast SgMg performs at nearly 300 FPS, which is about **5**× faster than SgMg.

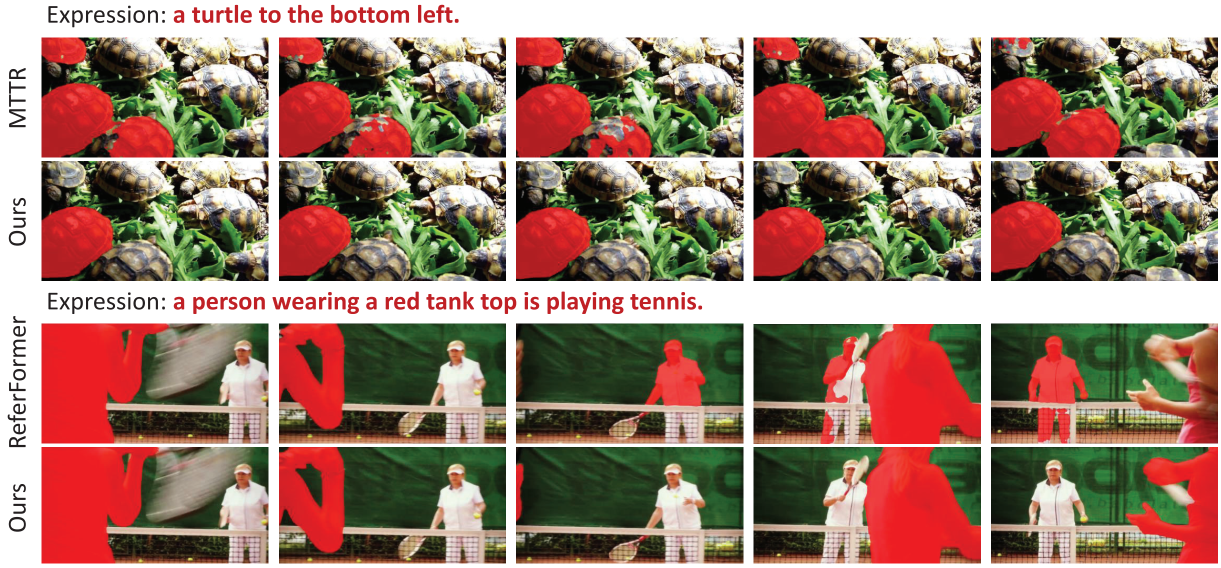

Figure 5. Qualitative comparison of our method with others.

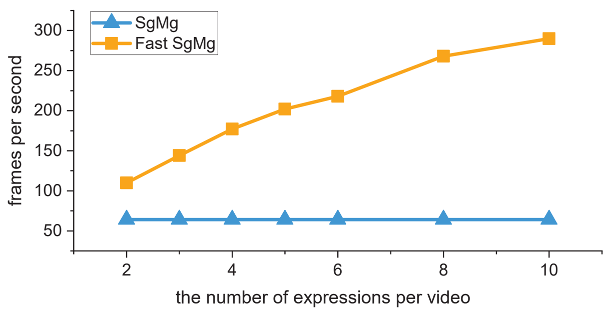

Figure 6. Efficiency analysis of SgMg and Fast SgMg for videos with different numbers of expressions on Ref-YouTube-VOS.

4.8. Qualitative Results

In Fig. 5, we show qualitative comparison with ReferFormer [57] and MTTR [4]. SgMg can handle different objects of the same category or with the same behavior.

4.9. Feature Visualization of SCF

In Fig. 7, we visualize the vision-language features extracted by our SCF in comparison to the cross-attention used in [57]. The features extracted by SCF exhibit superior grounding ability in locating target objects, resulting in better performance for SgMg.

5. Conclusion

We discovered the feature drift issue in current referring video object segmentation (R-VOS) methods, which negatively affects the segmentation kernels. We presented SgMg, a novel segment-and-optimize approach for R-VOS that avoids the drift issue and optimizes masks with visual

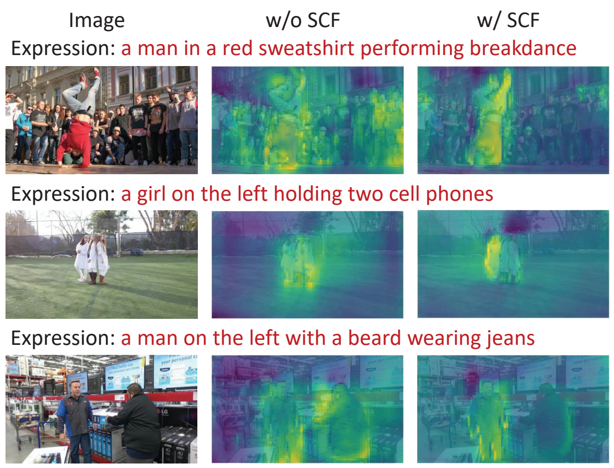

Figure 7. Visualization of the vision-language features extracted w/o and w/ our SCF.

details. We also provided a new perspective to encourage vision-language global interactions in the spectral domain with Spectrum-guided Cross-modal Fusion. Additionally, we proposed the multi-object R-VOS paradigm by extending SgMg with multi-instance fusion and decoupling. Finally, we evaluated our models on four video benchmarks and demonstrated that our approach achieves state-of-the-art performance on all four datasets.

Acknowledgment. This research was supported by the Australian Research Council Industrial Transformation Research Hub IH180100002. Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 2

[2] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications*, pages 1–20, 2022. 1

[3] Glenn D Bergland. A guided tour of the fast fourier transform. *IEEE spectrum*, 6(7):41–52, 1969. 3

[4] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4985–4995, 2022. 1, 2, 3, 4, 5, 6, 8

[5] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 2

[6] Weidong Chen, Dexiang Hong, Yuankai Qi, Zhenjun Han, Shuhui Wang, Laiyun Qing, Qingming Huang, and Guorong Li. Multi-attention network for compressed video referring object segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4416–4425, 2022. 1, 5, 6

[7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 34:11781–11794, 2021. 1, 2

[8] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 3

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7

[10] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973, 2022. 2, 6

[11] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, page 7, 2021. 2, 5, 6

[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2

[13] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021. 2

[14] Guang Feng, Lihe Zhang, Zhiwei Hu, and Huchuan Lu. Deeply interleaved two-stream encoder for referring video segmentation. *arXiv preprint arXiv:2203.15969*, 2022. 2

[15] Mingtao Feng, Haoran Hou, Liang Zhang, Yulan Guo, Hongshan Yu, Yaonan Wang, and Ajmal Mian. Exploring hierarchical spatial layout cues for 3d point cloud based scene graph prediction. *IEEE Transactions on Multimedia*, 2023. 1

[16] Zhiheng Fu, Siyu Hong, Mengyi Liu, Hamid Laga, Mohammed Bennamoun, Farid Boussaid, and Yulan Guo. Multi-stage information diffusion for joint depth and surface normal estimation. *Pattern Recognition*, 141:109660, 2023. 1

[17] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018. 2, 5, 6

[18] Lorenzo Giambagli, Lorenzo Buffoni, Timoteo Carletti, Walter Nocentini, and Duccio Fanelli. Machine learning in spectral domain. *Nature communications*, 12(1):1–9, 2021. 3

[19] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 6

[20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 2, 5

[21] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018. 1, 2, 5

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[23] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *AAAI Conference on Artificial Intelligence*, 2022. 1, 5, 6

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[25] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 7

[26] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 5

[27] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Yan Lu, and Bhiksha Raj. Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency. *arXiv preprint arXiv:2207.01203*, 2022. 1

[28] Zizhang Li, Mengmeng Wang, Jianbiao Mei, and Yong Liu. Mail: A unified mask-image-language trimodal network for referring image segmentation. *arXiv preprint arXiv:2111.10747*, 2021. 7

[29] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021. 2, 6

[30] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 2

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[32] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 6

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[34] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *European Conference on Computer Vision*, pages 648–665. Springer, 2022. 3

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[37] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 2

[38] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3623–3632, 2019. 2

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6, 7

[40] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 3

[41] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Region aware video object segmentation with deep motion modeling. *arXiv preprint arXiv:2207.10258*, 2022. 1

[42] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Self-supervised video object segmentation by motion-aware mask propagation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2022. 2

[43] Bo Miao, Liguang Zhou, Ajmal Mian, Tin Lun Lam, and Yangsheng Xu. Object-to-scene: Learning to transfer object knowledge to indoor scene recognition. In *Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2021. 1

[44] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proc. Int. Conf. Comput. Vis. (ICCV)*, pages 9226–9235, 2019. 1, 2

[45] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 5

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[47] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15455–15464, 2021. 2

[48] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[49] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223. Springer, 2020. 1, 2, 5, 6

[50] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 1, 2, 3, 4

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 7

[53] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast

end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 2

[54] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng 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*, pages 3939–3948, 2019. 6

[55] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 7

[56] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing 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*, pages 4996–5005, 2022. 1, 5, 6

[57] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[58] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. *arXiv preprint arXiv:2308.02874*, 2023. 1

[59] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019. 4

[60] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, pages 264–274. Springer, 2019. 4

[61] Zhao Yang, Yansong Tang, Luca Bertinetto, Hengshuang Zhao, and Philip HS Torr. Hierarchical interaction network for video object segmentation from referring expressions. In *BMVC*, 2021. 1, 2

[62] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 7

[63] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 5, 6

[64] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[65] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 6, 7

[66] Z. Yang and Y, Wei and Y. Yang. Associating objects with transformers for video object segmentation. *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 34:2491–2502, 2021. 2

[67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[68] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2305.18010*, 2023. 2

[69] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. *arXiv preprint arXiv:2205.00823*, 2022. 2

[70] Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, and Yang You. Modeling motion with multi-modal features for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2022. 2

[71] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2

[72] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3, 5, 6