# 3rd Place Solution for The 5th Large-scale Video Object Segmentation: Challenge——Track 3: Referring Video Object Segmentation

Bo Miao[1], Zijie Wu[2], Mohammed Bennamoun[1], Yongsheng Gao[3], Ajmal Mian[1]
[1]The University of Western Australia [2]Hunan University [3]Griffith University

## Abstract

*Referring video object segmentation (R-VOS) aims to segment objects of interest in video, referred to by linguistic expressions. In this report, we present our solution for the 5th Large-scale Video Object Segmentation Challenge, built upon SgMg [19]. Unlike previous R-VOS techniques that follow a decode-and-segment paradigm, SgMg adopts an efficient segment-and-refine paradigm to address the feature drift issue and achieve top-ranked performance. Without bells and whistles,* e.g.*, joint training and test-time augmentations, our solution achieves 60.0 $\mathcal{J}\&\mathcal{F}$ on the test split of Ref-YouTube-VOS and ranked $3^{rd}$ place in Track 3 (Referring Video Object Segmentation) of the 5th Large-scale Video Object Segmentation Challenge. Moreover, we outperform existing state-of-the-art competitors in a fair comparison. Code is available at https://github.com/bo-miao/SgMg.*

## 1. Introduction

Referring video object segmentation (R-VOS) is an emerging video task that aims to segment target objects in video, referred to by linguistic expressions. It benefits a wide range of applications such as video surveillance. Unlike semi-supervised video object segmentation [28, 3, 17, 18], which benefit from provided ground truth for the first frame, R-VOS is more challenging due to the requirement for cross-modal understanding.

Recent R-VOS techniques employ attention-based transformers to capture long-range dependencies and handle multimodal features, achieving promising performance. Based on the diverse object queries, conditional kernel [22] is then introduced [1, 25] to dynamically identify target objects within videos. These methods follow a decode-and-segment paradigm, where kernels are extracted from encoded features to segment decoded features. Despite their promising performance, this paradigm suffers from feature drift issues which hampers the effectiveness of the kernels.

In this report, we present our solution for the R-VOS challenge, which is entirely based on SgMg [19]. SgMg employs the conditional kernel to directly segment its fully perceived encoded features to generate mask priors, preventing the feature drift and its adverse effects. The priors are then refined using visual details to generate fine-grained masks. We conduct experiments on Ref-YouTube-VOS to validate the effectiveness of SgMg. Even without using joint training and test-time augmentations, SgMg achieves **60.0** $\mathcal{J}\&\mathcal{F}$ on the Ref-YouTube-VOS test split, and ranked $3^{rd}$ place in Track 3 (Referring Video Object Segmentation) of the 5th Large-scale Video Object Segmentation Challenge.

## 2. Related Works

**Referring Video Object Segmentation.** Current methods utilize multimodal interactions to equip visual features with correlated linguistic information for R-VOS. [21] proposes a unified R-VOS framework that conducts iterative segmentation using linguistic and temporal features. [10] establishes object relations and tracklets for sequence-level segmentation. [26, 5] perform hierarchical cross-modal fusion to improve feature representations. [12, 9] conducts progressive segmentation that perceives object candidates and then finds the optimal match.

With the advance of transformers [23], MTTR [1] introduces an end-to-end network with conditional kernels [22] for dynamic segmentation and achieves impressive performance. ReferFormer [25] further proposes language-guided conditional kernels, which are object-specific, to boost performance. However, the decode-and-segment paradigm within their methods leads to feature drift issues, making the network sub-optimal. SgMg [19] proposes a segment-and-optimize paradigm to address the drift problem, achieving state-of-the-art performance with efficient inference time. In this challenge, we employ SgMg to evaluate the test split of Ref-YouTube-VOS.

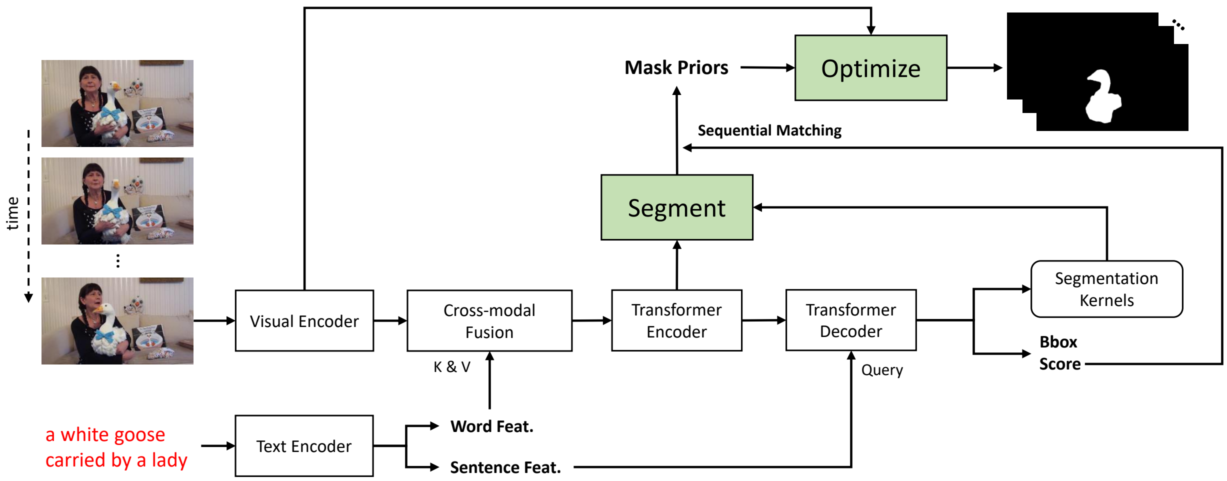


Figure 1. The overall framework of SgMg, simplified from [19]. Given a video sequence and a language description, the cross-modal fusion enhances visual features using linguistic information. Language-guided queries associate vision-language features to generate segmentation kernels and predict mask priors. The optimization recovers visual details for the priors and generates fine-grained results.

# 3. Method

This section presents the SgMg approach [19], including the cross-modal fusion and the segment-and-optimize paradigm. The overall framework is shown in Fig. 1. In this report, we adopt VideoSwin [14] and RoBERTa [13] as visual and text encoders respectively. More details can be found in [19].

## 3.1. Cross-modal Fusion

SgMg [19] includes a spectrum-guided cross-modal fusion to improve multimodal representations. The module conducts **spectrum augmentation** with adaptive Gaussian smoothed filters to enhance features before and after cross-attention between visual and textual representations. Given the input feature map $\mathbf{F}$, spectrum augmentation (SA) is computed as:

$$\mathrm{SA}(\mathbf{F}, K) = \mathbf{F} + \Phi_{\mathrm{IFFT}}(\mathrm{Proj}(\sigma(K, \mathbf{F}) \odot \Phi_{\mathrm{FFT}}(\mathbf{F}))) \quad (1)$$

where $\odot$ denotes low-pass filtering through adaptive Gaussian smoothed filters $\sigma(K, \mathbf{F})$, a 2D Gaussian map generated based on the bandwidth $K$ and scaled by a parameter predicted from $\mathbf{F}$. The point-wise spectral operations in SA promote global interactions and thus enhance feature representations. In summary, the spectrum-guided cross-modal fusion can be represented as:

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

where $\mathbf{F}_v$ and $\mathbf{F}_w$ are visual and textual features.

## 3.2. Segment-and-Optimize Paradigm

The segment-and-optimize paradigm proposed by SgMg [19] conditionally segments encoded features to predict (patch) mask priors and performs multi-granularity optimization to recover visual details.

For **conditional segmentation**, language-guided object queries $Q$ interact with vision-language features $\mathbf{F}_{vl}$ to predict kernels,

$$\mathrm{Kernel}(Q, \mathbf{F}_{vl}) = \Phi(\mathrm{Proj}(\mathrm{Att}(Q, \mathbf{F}_{vl}))) \quad (3)$$

where $\Phi$ represents the parameterization operation to generate two point-wise convolutions. The kernels convolve on $\mathbf{F}_{vl}$ to predict mask priors.

For **multi-granularity optimization**, SgMg reuses visual features with spatial strides of $\{4{,}8\}$ to predict residual maps of mask priors, progressively recovering visual details to efficiently generate fine-grained masks.

## 3.3. Sequential Matching and Loss Functions

We perform instance matching with five language-guided object queries. Each query predicts a bounding box $\mathbf{B}$, a score $\mathbf{S}$ indicating mask quality, and conditional kernels generating mask priors $\mathbf{M}_P$. The Hungarian algorithm [6] is adopted to find the best result (query), and the multi-granularity optimizer refines the optimal $\mathbf{M}_P$ to produce full-resolution mask $\mathbf{M}$.

To supervise the model, we use Dice loss [8] and Focal loss [11] for masks, Focal loss [11] for scores, and L1 and GIoU [20] loss for bounding boxes:

$$\mathcal{L}_{train} = \lambda_m(\mathcal{L}_{\mathbf{M}_P} + \mathcal{L}_{\mathbf{M}}) + \lambda_b\mathcal{L}_{\mathbf{B}} + \lambda_s\mathcal{L}_{\mathbf{S}} \quad (4)$$

where $\mathcal{L}$ and $\lambda$ are the loss term and weight.

# 4. Implementation Details

Following [25, 19], we first pre-train our model on RefCOCO/+/g [16, 27] and then fine-tune it on the training set of Ref-YouTube-VOS [21]. The model is trained using

| Team | Overall | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| Robertluo | 70.0 | 68.0 | 72.0 |
| beter | 66.0 | 64.0 | 68.0 |
| **Ours** | **60.0** | **59.0** | **62.0** |
| MahouShoujo | 60.0 | 58.0 | 61.0 |

Table 1. The leaderboard of the R-VOS challenge.

AdamW [15] optimizer for 12 epochs in pre-training and 6 epochs in main training. During pre-training, we set the initial learning rates of 2.5e-6, 1.25e-5, and 2.5e-5 for the text encoder, visual encoder, and the rest components, respectively. The pre-training uses a single frame, with the learning rates decayed by a factor of 10 at the 8th and 10th epochs. In the main training, we freeze the text encoder, and the initial learning rates of 2.5e-5 and 5e-5 are adopted for the visual encoder and the rest. The learning rates are divided by 10 at the 3rd and 5th epochs.

The model is trained on 2 RTX 3090 GPUs with 5 randomly selected frames per clip, all resized to the longest side of 640 pixels. The coefficients for different loss terms $\lambda_{dice}$, $\lambda_{focal}$, $\lambda_{L1}$, $\lambda_{giou}$ are set to 5, 2, 5, and 2. The data augmentation comprises random resize, random crop, random horizontal flip, and photometric distortion.

## 5. The 5th Large-scale Video Object Segmentation Challenge

Our result ranked 3rd in the 5th YouTube-RVOS Challenge, without using techniques like joint training or test-time augmentations. As shown in Table 1, we achieved an overall accuracy of 60.0 on the Ref-YouTube-VOS 2023 test set. For a fair comparison with previous benchmarks, we conducted the evaluation under identical settings on the validation split of Ref-YouTube-VOS. As shown in Table 2, we achieved 58.9 $\mathcal{J}\&\mathcal{F}$, outperforming the nearest competitor by 2.9% points.

## 6. Conclusion

We employed the efficient SgMg [19] for the R-VOS challenge. SgMg follows a segment-and-optimize paradigm to address feature drift issues exist in prior methods, while its spectrum-guided cross-modal fusion enhances multi-modal feature representations. Without bells and whistles, Our solution ranked $3^{rd}$ in Track 3 (Referring Video Object Segmentation) of the 5th Large-scale Video Object Segmentation Challenge and remarkably outperforms previous benchmarks on the validation split of Ref-YouTube-VOS. We hope SgMg will serve as a solid baseline for R-VOS and benefit other approaches encountering the drift issue.

**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.

| Method | Ref-YouTube-VOS | | |
|---|---|---|---|
| | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
| URVOS [21] | 47.2 | 45.3 | 49.2 |
| CMPC-V [12] | 47.5 | 45.6 | 49.3 |
| PMINet [5] | 53.0 | 51.5 | 54.5 |
| YOFO [7] | 48.6 | 47.5 | 49.7 |
| LBDT [4] | 49.4 | 48.2 | 50.6 |
| MLRL [24] | 49.7 | 48.4 | 51.0 |
| MTTR [1] | 55.3 | 54.0 | 56.6 |
| MANet [2] | 55.6 | 54.8 | 56.5 |
| ReferFormer [25] | 56.0 | 54.8 | 57.3 |
| **Ours** | **58.9** | **57.7** | **60.0** |

Table 2. Comparison to state-of-the-art methods on the validation split of Ref-YouTube-VOS, excerpted from [19].

## References

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

[2] 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 *ACM MM*, pages 4416–4425, 2022. 3

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 34:11781–11794, 2021. 1

[4] 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. 3

[5] 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. 1, 3

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

[7] 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*, 2022. 3

[8] 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. 2

[9] 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. 1

[10] 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. 1

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 2

[12] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *TPAMI*, 2021. 1, 3

[13] 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. 2

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*. 2

[17] 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

[18] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Self-supervised video object segmentation by motion-aware mask propagation. In *ICME*, 2022. 1

[19] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. *arXiv preprint arXiv:2307.13537*, 2023. 1, 2, 3

[20] 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 *CVPR*, pages 658–666, 2019. 2

[21] 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, 3

[22] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, pages 282–298. Springer, 2020. 1

[23] 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

[24] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *CVPR*, pages 4996–5005, 2022. 3

[25] 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

[26] 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

[27] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2

[28] 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. 1