

ReferDINO-Plus: 2nd Solution for 4th PVUW MeViS Challenge at CVPR 2025

Tianming Liang Haichao Jiang Wei-Shi Zheng Jian-Fang Hu*

Sun Yat-sen University

Abstract

Referring video object segmentation (RVOS) aims to segment target objects throughout a video based on a text description. This task has attracted increasing attention in the field of computer vision due to its promising applications in video editing and human-agent interaction. Recently, ReferDINO showcases promising performance in this task by adapting the object-level vision-language knowledge from pretrained foundational image models. In this report, we extend its capabilities by incorporating SAM2 to enhance mask quality and object consistency. To effectively balance performance between single-object and multi-object scenarios, we introduce a conditional mask fusion strategy that adaptively combines masks from ReferDINO and SAM2. Our solution, termed ReferDINO-Plus, achieves 60.43 $\mathcal{J}\&\mathcal{F}$ on MeViS test set, securing 2nd place in the MeViS PVUW challenge at CVPR 2025. The code is available at: <https://github.com/iSEE-Laboratory/ReferDINO-Plus>.

1. Introduction

Referring video object segmentation (RVOS) aims to segment target objects throughout a video based on a text description. This task bridges the gap between vision-language understanding and pixel-level video analysis, offering significant value for many down-stream applications, such as video editing and human-agent interaction systems.

Previous RVOS datasets like Refer-Youtube-VOS [21] and Ref-DAVIS17 [13] focused on segmenting salient video objects described by static attributes (e.g., color and shape) or simple spatial relationships, overlooking the complex, dynamic properties in real-world scenarios. To encourage more efforts towards challenging yet practical pixel-level video understanding, the 4th PVUW workshop at CVPR 2025 presents a challenging RVOS benchmark MeViS [7] for competition. Different from previous RVOS datasets, MeViS [7] focuses on the understanding of temporal motion in the RVOS task. In MeViS, the videos often contain

multiple objects with similar static appearances but different motion attributes, and the object descriptions in MeViS mainly focus on motion and temporal expressions. In addition, MeViS includes numerous multi-object expressions, allowing for the referral of an unlimited number of target objects in the video. These features make MeViS more challenging and reflective of real-world scenarios. To overcome these challenges, a strong cross-modal spatiotemporal capability is necessary to understand the motion properties in the videos and descriptions.

Early works [1] in RVOS tend to directly apply the referring image segmentation methods [6, 24] to RVOS. However, this manner ignores the temporal information and often result in inconsistent object prediction. Afterwards, MTTR [2] introduced the DETR paradigm [3] into RVOS. Building on this, ReferFormer [22] proposed to generate queries directly from the text description. Subsequent works [10, 17, 19, 25] have focused on modular improvements to enhance cross-frame consistency and temporal understanding. Despite these efforts, current RVOS models [11, 17, 23] still struggle with insufficient vision-language understanding, often failing to handle complicated object descriptions, especially involving composite appearance, location and attributes. Recently, ReferDINO [15] was proposed to address this limitation by leveraging the pretrained vision-language knowledge from the foundational visual-grounding model GroundingDINO [16]. To enable end-to-end adaptation on RVOS data, ReferDINO incorporates a cross-modal temporal enhancer and a well-designed mask decoder. Combining these components, ReferDINO achieves state-of-the-art performance across various RVOS benchmarks.

Our solution, termed **ReferDINO-Plus**, is a two-stage strategy built upon ReferDINO and SAM2 [20]. Specifically, in the first stage, we employ ReferDINO to perform cross-modal object identification and spatiotemporal dense reasoning. Given a video and an object description, ReferDINO generates masks and binary scores for each candidate target. However, due to the lack of training on large-scale segmentation data, the mask quality may be unsatisfactory. Therefore, in the second stage, we apply SAM2 for mask refinement and augmentation, regarding the frame and mask

*Corresponding author.

with the highest binary score as the prompts. After this two-stage process, we can obtain two series of masks—one from ReferDINO and the other from SAM2. Intuitively, the masks from SAM2 are more reliable and stable. However, we observe that SAM2 tends to degenerate the multi-object mask into a single-object mask, resulting in performance degradation in multi-object scenarios. To address this issue, we design a Conditional Mask Fusion (CMF) strategy. For single-object cases, we output only the masks from SAM2; for multi-object cases, we combine both the masks from ReferDINO and SAM2. However, it remains challenging to determine whether an expression involves multiple objects. In our experiment, we define it as a multi-object case if the mask area of SAM2 is less than $2/3$ of ReferDINO’s. Our solution is straight-forward yet effective. Without further finetuning with additional pseudo labels on validation/test data [8], our solution achieves 55.27 $\mathcal{J}\&\mathcal{F}$ on MeViS validation set, and 60.43 $\mathcal{J}\&\mathcal{F}$ on MeViS test set, securing the final ranking of 2nd in the MeViS Track at CVPR 2025 PVUW challenge.

2. Related Works

2.1. Referring Video Object Segmentation

RVOS [7, 9, 21] aims to segment objects throughout the video based on text descriptions. Some works [1] attempt to directly apply the referring image segmentation methods [6, 24] to RVOS. However, this manner is unable to capture temporal information and often result in inconsistent object prediction. MTTR [2] firstly introduces the DETR paradigm [3] into RVOS. Furthermore, ReferFormer [22] proposes to produce the queries from the text description. On the top of this pipeline, follow-up works [10, 17, 19, 25] focus on modular improvements to improve cross-frame consistency and temporal understanding. For example, SOC [17] aggregates video content and textual guidance with a semantic integration module for unified temporal modeling and cross-modal alignment. DsHmp [11] decouples video-level referring expression understanding into static and motion perception, with a customized module to enhance temporal comprehension. Despite notable progress on specific datasets, these models are limited by insufficient vision-language understanding, and often struggle in unseen objects or scenarios. Recently, ReferDINO [15] overcomes this limitation by leveraging the pretrained vision-language knowledge from GroundingDINO [16], and extending the capabilities of dense perception and spatio-temporal reasoning by integrating an effective mask decoder and a temporal enhancer.

2.2. Semi-supervised Video Object Segmentation

The conventional semi-supervised video object segmentation aims to propagate the ground-truth object masks from

a given frame throughout the video. Many existing approaches [4, 5, 20] employ a memory mechanism to store the past features for tracking and segmenting on future frames. Early deep learning-based methods mainly employed online adaptation strategies, where models were fine-tuned either on the initial frame or all frames to specialize in the target object. To reduce the computation overheads, many works focus on offline training conditioned solely on the first frame or incorporating temporal dependencies from preceding frames. With the development of the strong Segment Anything Model (SAM) [14], many efforts attempt to combine SAM with video trackers based on masks to perform segmentation throughout the video. However, this manner is ineffective since the combination is not end-to-end differentiable. To address this limitation, SAM2 [20] was proposed, aiming to accommodate the specific demand of semi-supervised video object segmentation. SAM2 showcases strong performance in object tracking and segmentation, and has attracted more and more attention in the computer vision field.

3. ReferDINO-Plus

The overall framework of our solution **ReferDINO-Plus** is presented in Figure 1. For each video-description pair, we input it into ReferDINO to derive the object masks and the corresponding scores across the frames. Then, we select the mask with highest score as the prompt of SAM2, producing refined masks. Finally, we fuse the two series of masks through the conditional mask fusion strategy, to generate the final masks on each frame.

3.1. Cross-modal Reasoning with ReferDINO

ReferDINO [15] is a strong RVOS model that inherits object-level vision-language knowledge from GroundingDINO [16], and is further endowed with pixel-level dense prediction and cross-modal spatiotemporal reasoning. Formally, given a video clip of T frames and a text description, ReferDINO performs cross-modal reasoning and segmentation, deriving a mask sequence $\{M_r^t\}_{t=1}^T$ and the corresponding scores $\{S_r^t\}_{t=1}^T$ throughout the video. Following the practice in previous works [7, 11, 15], we combine the multiple object masks with scores higher than a preset threshold σ to handle the multi-object cases.

3.2. Mask Refinement with SAM2

SAM2 [20] is a strong prompt-based segmentation model, which can efficiently produce high-quality object masks throughout the video based on the given prompts, in form of clicks, bounding boxes and masks. In this work, we utilize SAM2 to further refine the mask quality and object consistency of ReferDINO. Specifically, after obtaining the masks and corresponding scores across the frames, we select the mask with the maximum score as the prompt. Based on the

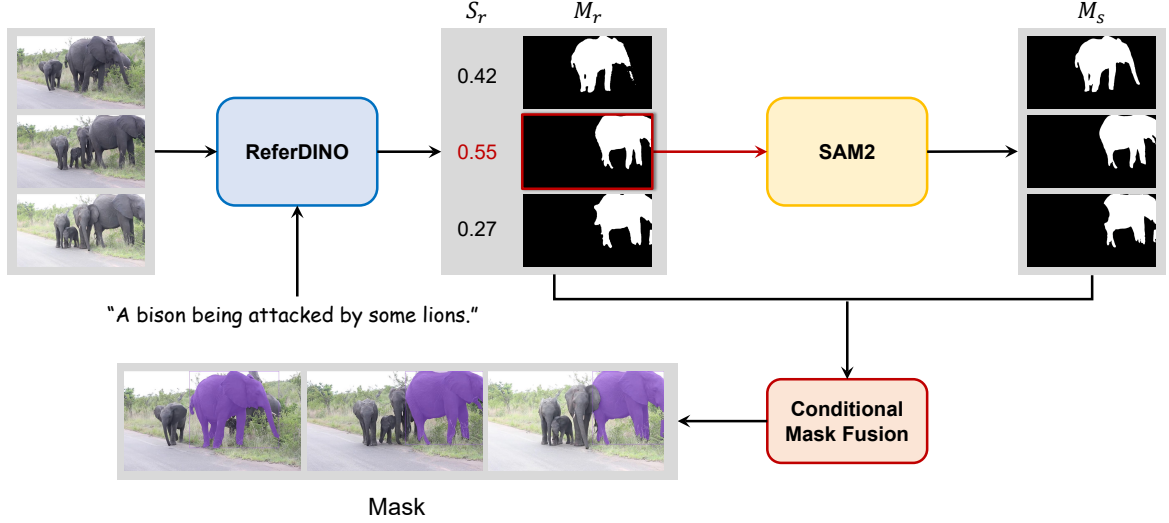


Figure 1. Overview of our solution **ReferDINO-Plus**. For each video-description pair, we input it into ReferDINO to derive the object masks M_r and the corresponding scores S_r across the frames. Then, we select the mask with highest score as the prompt of SAM2, producing refined masks M_s . Finally, we fuse the two series of masks through the *conditional mask fusion* strategy. Best view in color.

prompt frame and mask, SAM2 produces a refined mask sequence $\{M_s^t\}_{t=1}^T$ throughout the video.

3.3. Conditional Mask Fusion

Although the masks from SAM2 are more reliable and stable, we observe that SAM2’s overall performance on MeViS is significantly weaker than that of ReferDINO. In our experiments, we identify the main reason as that, for multi-object mask prompts, SAM2 tends to degenerate them into single-object masks, leading to a substantial target loss in subsequent frames. To address this issue, we design a *Conditional Mask Fusion* (CMF) principle: for single-object cases, we output only the masks from SAM2; for multi-object cases, we combine both the masks from ReferDINO and SAM2.

However, it remains challenging to determine whether an expression involves multiple objects. In our solution, we define it as a multi-object case if the mask area of SAM2 is less than $2/3$ of ReferDINO’s. Formally, this process can be described as follows:

$$M = \begin{cases} M_s + M_r & \text{if } \mathcal{A}(M_s) < \frac{2}{3}\mathcal{A}(M_r) \\ M_s & \text{otherwise} \end{cases} \quad (1)$$

where $\mathcal{A}(\cdot)$ indicates the mask area. Note that our CMF is conducted individually on each frame, which empirically achieves better performance.

4. Experiment

4.1. Dataset and Metrics

MeViS [7] is a large RVOS dataset comprising 2K videos and 28K text descriptions. In this competition, the provided test set includes 100 videos and 1,456 language descriptions. These descriptions may correspond to a single object, multiple objects, or even non-objects within the videos, making the dataset significantly challenging. We employ region similarity \mathcal{J} (average IoU), contour accuracy F (mean boundary similarity), and their average $\mathcal{J}\&\mathcal{F}$ as the evaluation metrics.

4.2. Implementation Details

We pretrain ReferDINO [15] on the referring image segmentation datasets RefCOCO/+g [12, 18] at first, and then train with the combination of Refer-Youtube-VOS [21] and Ref-DAVIS17 [13]. Finally, we finetune it on the training set of MeViS. For ReferDINO, we use the MM-GroundingDINO-SwinB [26] as the backbone. For SAM2, we use the Sam2.1_Hiera_Large as the backbone. We set the threshold $\sigma = 0.275$, other hyper-parameters are the same as the original ReferDINO. Unlike the solutions [8] in previous challenges, we do not use additional pseudo labels on the validation or test data to for further finetuning.

4.3. Competition Results

As shown in Table 1, our solution achieves 60.43 $\mathcal{J}\&\mathcal{F}$ on MeViS test set, securing the final ranking of 2nd in the MeViS Track at CVPR 2025 PVUW challenge.

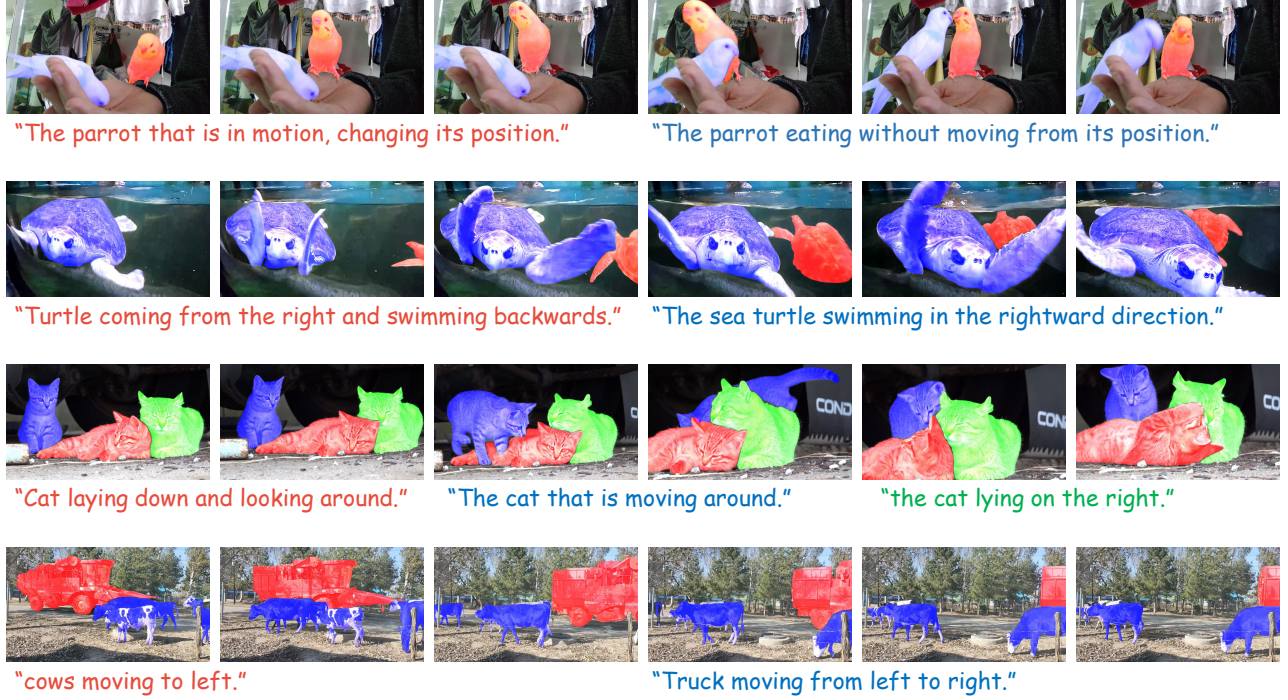


Figure 2. Visualization results of our solution on MeViS test set.

Team	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
MVP-Lab	61.98	58.83	65.14
ReferDINO-Plus	60.43	56.79	64.07
HarborY	56.26	52.68	59.84
Pengsong	55.91	53.06	58.76
ssam2s	55.16	52.00	58.33
strong_kimchi	55.02	51.78	58.27

Table 1. The leaderboard of the MeViS test set.

Method	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
ReferDINO	51.67	47.94	55.40
+SAM2	52.54	49.18	55.90
+SAM2+CMF _V	54.82	51.39	58.24
+SAM2+CMF	55.27	51.80	58.75

Table 2. Ablation studies on the MeViS validation set.

4.4. Ablation Studies

We conduct ablation studies on MeViS validation set to explore the effects of individual components in our solution. As shown in Table 2, the refinement of SAM2 improves 3.15% $\mathcal{J}\&\mathcal{F}$. When we perform CMF on the entire video (termed CMF_V), the result can be improved by 2.28% $\mathcal{J}\&\mathcal{F}$. When performing CMF on individual frames, the

result can be further improved by 0.45%. These results demonstrate the effectiveness of our components.

4.5. Visualization

In Figure 2, we present several visualization results of ReferDINO-Plus on MeViS test set. It shows that our method can effectively segment the targets based on the corresponding text descriptions throughout the videos. These results demonstrate the accurate and high-quality masks generated from our ReferDINO-Plus. In the 4th line of Figure 2, we further show a case of multi-object referring, which demonstrates the effectiveness of our conditional mask fusion strategy.

5. Conclusion

In this work, we propose ReferDINO-Plus to address the problem of motion expression guided video object segmentation. This is a two-stage strategy. In the first stage, it employs ReferDINO to perform cross-modal object identification and spatiotemporal dense reasoning. In the second stage, it integrates SAM2 for mask refinement and object tracking. To address the multi-object collapse problem, we further design a conditional mask fusion strategy for post segmentation ensemble. Without further finetuning with additional pseudo labels on validation/test data, our solution achieves the 2nd place in PVUW MeViS challenge at CVPR 2025.

References

- [1] M Bellver, C Ventura, C Silberer, I Kazakos, J Torres, and X Giro-i Nieto. Refvos: a closer look at referring expressions for video object segmentation (2020), 2010. 1, 2
- [2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multi-modal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 2
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2
- [4] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2
- [5] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 2
- [6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 1, 2
- [7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 1, 2, 3
- [8] Hao Fang, Feiyu Pan, Xiankai Lu, Wei Zhang, and Runmin Cong. Uninext-cutie: The 1st solution for lsvos challenge rvos track. *arXiv preprint arXiv:2408.10129*, 2024. 2, 3
- [9] Kirill Gavriluk, 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
- [10] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Htm1: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13414–13423, 2023. 1, 2
- [11] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13332–13341, 2024. 1, 2
- [12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3
- [13] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1, 3
- [14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2
- [15] Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. *arXiv preprint arXiv:2501.14607*, 2025. 1, 2, 3
- [16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. 1, 2
- [17] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: semantic-assisted object cluster for referring video object segmentation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26425–26437, 2023. 1, 2
- [18] 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. 3
- [19] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 1, 2
- [20] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2
- [21] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1, 2, 3
- [22] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 2
- [23] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongqiang He, and Peng Gao. Referred by multi-modality: A unified tem-

poral transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6449–6457, 2024. [1](#)

- [24] 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. [1](#), [2](#)
- [25] Linfeng Yuan, Miaoqing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14001–14010, 2024. [1](#), [2](#)
- [26] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. [3](#)