WHO DOES WHAT TO WHOM? ROLE BINDING FOR INTERACTION-AWARE VIDEO SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interaction-aware referring video object segmentation (InterRVOS) asks not only which objects are referred to, but who does what to whom: the model must segment the Actor and Target of a described interaction according to their asymmetric semantic roles. Unlike conventional referring segmentation, these roles are not intrinsic object identities but are defined by the interaction between participants. Existing methods use separate Actor and Target tokens while leaving this interaction implicit, which can lead to ambiguous or relationally inconsistent role predictions. We instead formulate InterRVOS as relation-conditioned role binding, where the interaction serves as an explicit intermediate state that connects multimodal understanding to role-specific segmentation. Based on this view, we introduce R^{3}Refer, which first learns an input-dependent, visually grounded interaction representation. It then performs directed role reasoning, where the interaction conditions Actor prediction and the resulting Actor representation further guides Target reasoning. Finally, triadic refinement jointly coordinates the Interaction, Actor, and Target representations before mask decoding. Experiments on InterRVOS-127K show consistent improvements in both Actor and Target segmentation, while evaluations on conventional RVOS benchmarks demonstrate the broader effectiveness of relation-aware role modeling.

1 INTRODUCTION

Interaction-aware referring video object segmentation (InterRVOS) Jin et al. (2026) asks a model to determine *who does what to whom*: given a natural-language expression describing an interaction, it must separately segment the Actor performing the interaction and the Target toward which it is directed. This requirement fundamentally differs from standard referring video object segmentation (RVOS) Khoreva et al. (2018); Seo et al. (2020); Ding et al. (2023); Botach et al. (2022); Wu et al. (2022), where the primary goal is to identify the object referred to by an expression. In InterRVOS, Actor and Target are not intrinsic properties of individual objects. The same participant can occupy different roles under different interaction descriptions, meaning that its role is determined by how it relates to other participants. InterRVOS therefore requires more than object localization: it requires *binding visual entities to the asymmetric roles defined by an interaction*.

Recent multimodal large language model (MLLM)-based segmentation methods Lai et al. (2024); Yan et al. (2024a); Bai et al. (2024); Yuan et al. (2025); Munasinghe et al. (2025); Wei et al. (2025); Liu et al. (2026); Gong et al. (2025); Hong & Do (2026); Zhang et al. (2026) connect language understanding to mask prediction through dedicated segmentation tokens. ReVIOSa Jin et al. (2026), introduced with InterRVOS, extends this paradigm by assigning separate Actor and Target tokens for role-specific segmentation. This makes the two output roles explicit within a unified model. However, it leaves implicit the very relation that determines how visual entities should be assigned to those roles. Actor and Target representations are still largely formed from a shared multimodal context, requiring each representation to recover the relevant relational cues on its own.

As illustrated in Fig. 1, existing methods explicitly represent the Actor and Target outputs but leave implicit the interaction that defines their roles. We argue that this creates a structural mismatch between the task and its representation. The outputs are relationally defined, yet the relation itself is not explicitly represented during role formation. Consequently, Actor and Target can degenerate into two loosely coupled localization problems, producing predictions that are individually plausi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ble but incompatible with the described interaction. More importantly, this formulation underuses the asymmetry inherent in directed interactions. Once an Actor hypothesis is established under a particular interaction, it provides strong context for determining which participant can occupy the complementary Target role. Thus, Actor and Target should not be inferred as parallel and independent outputs, but as interdependent role assignments conditioned on a shared interaction.

This observation motivates us to formulate InterRVOS as *relation-conditioned role binding*. Rather than treating the interaction as information that must be rediscovered independently by each role representation, we introduce it as an explicit intermediate state between multimodal understanding and role-specific segmentation. Such a state should capture more than a discrete action label. The same semantic interaction can be instantiated through different spatial configurations, contacts, and motion patterns across videos. We therefore seek an input-dependent and visually grounded interaction representation that describes how the referred relation is instantiated in the current video. This idea is related to explicit relation modeling in video relation detection and scene graph generation Shang et al. (2017); Zheng et al. (2022); Cong et al. (2021); Yang et al. (2023), but serves a different purpose here: the relation is not merely an output to recognize, but an intermediate structure used to resolve which visual entities fill its semantic roles.

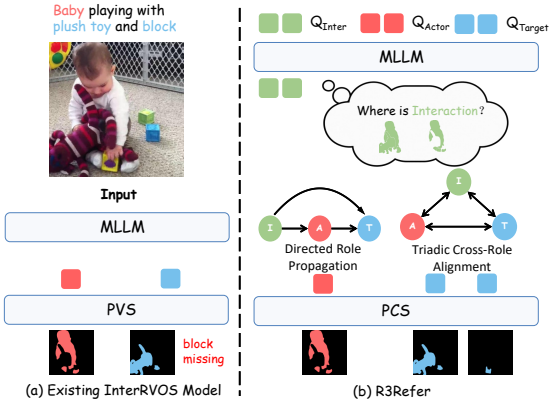

Figure 1: **Relation-conditioned role binding.** Existing methods predict Actor and Target while leaving their interaction implicit. We explicitly model the Interaction to guide and coordinate Actor–Target reasoning before segmentation.

Based on this formulation, we introduce R^{3}Refer, which organizes Actor–Target segmentation as a structured inference process over the interaction and its participants. First, *Dynamic Interaction Reasoning* (DIR) learns a visually grounded latent interaction state from the joint video-language context, establishing the relational context under which the two roles should be interpreted. Second, *Interaction-Anchored Role Calibration* (IARC) uses this state to resolve the asymmetric Actor–Target dependency. Its *Directed Role Propagation* (DRP) first conditions the Actor representation on the interaction and then allows the resulting Actor context, together with the interaction state, to guide Target reasoning. This ordering reflects the directed structure of the interaction: identifying one role narrows the plausible assignment of its complementary role. Finally, because interaction recognition and participant identification are mutually informative, *Triadic Cross-Role Alignment* (TCRA) allows the Interaction, Actor, and Target representations to exchange evidence and reconcile their predictions before mask decoding. In this way, R^{3}Refer follows a simple principle: first ground the relation, then bind its roles, and finally refine the relation and participants jointly.

We evaluate R^{3}Refer on InterRVOS-127K Jin et al. (2026) for interaction-aware Actor–Target segmentation and conventional RVOS, and further examine its generalization on standard RVOS benchmarks Seo et al. (2020); Khoreva et al. (2018); Ding et al. (2023); Yan et al. (2024a). R^{3}Refer consistently improves both Actor and Target segmentation on InterRVOS-127K, while remaining effective on conventional referring segmentation tasks. These results support the view that explicitly modeling the relation that defines the roles provides a more suitable representation for interaction-aware segmentation than predicting the two roles in isolation.

In summary, our contributions are threefold:

- We formulate InterRVOS as *relation-conditioned role binding*, explicitly modeling the interaction as the relational state that defines Actor and Target roles.
- We propose R^{3}Refer, which learns a visually grounded interaction representation and performs directed Actor–Target reasoning followed by triadic refinement for role-consistent segmentation.
- Extensive experiments on InterRVOS-127K demonstrate state-of-the-art Actor and Target segmentation, with strong generalization to conventional RVOS benchmarks.

2 Related Work

Referring Video Object Segmentation. Referring video object segmentation (RVOS) aims to segment objects specified by natural-language expressions throughout a video Gavrilyuk et al. (2018); Khoreva et al. (2018); Seo et al. (2020). Early studies mainly focused on direct object descriptions and extended image-level referring segmentation Hu et al. (2016); Kazemzadeh et al. (2014); Nagaraja et al. (2016) to temporally consistent video prediction Seo et al. (2020); Botach et al. (2022); Wu et al. (2022). Subsequent works introduced more challenging expressions that require motion understanding and temporal reasoning. MeViS Ding et al. (2023) considers motion-centric expressions involving single or multiple objects, while ReVOS Yan et al. (2024a) further incorporates visual attributes and temporal relations for reasoning-oriented segmentation. Recent MLLM-based approaches Lai et al. (2024); Yan et al. (2024a); Bai et al. (2024); Munasinghe et al. (2025); Yuan et al. (2025); Wei et al. (2025); Liu et al. (2026); Gong et al. (2025); Hong & Do (2026); Zhang et al. (2026) further couple language understanding with mask prediction through dedicated segmentation tokens or query representations, enabling more flexible reasoning over complex video expressions. InterRVOS Jin et al. (2026) further extends RVOS to interaction-aware scenarios by requiring separate segmentation of the Actor and Target involved in a described interaction. ReVIOSa Jin et al. (2026) employs role-specific segmentation tokens for Actor–Target prediction, while the interaction that determines these roles remains implicitly encoded, leaving each role representation to recover relational cues independently and limiting both relational reasoning and role-specific segmentation. Differently, we explicitly model the interaction as an intermediate relational representation for Actor-Target role reasoning.

Interaction Modeling for Video Understanding. Interaction modeling in videos typically represents relationships among entities through structured subject–relation–object formulations Shang et al. (2017); Tsai et al. (2019); Yang et al. (2024). Early video relation detection methods focused on detecting and tracking related object pairs Shang et al. (2017); Qian et al. (2019), while subsequent approaches modeled temporal dependencies to improve relation recognition across frames Tsai et al. (2019); Liu et al. (2020); Cong et al. (2021). VRDFormer Zheng et al. (2022) formulates video relation detection with relation queries, whereas DSG-DETR Feng et al. (2023) and TEMPURA Nag et al. (2023) further exploit temporal context to stabilize dynamic scene graph prediction. VrdONE Jiang et al. (2024) models subject–object interactions across multiple temporal scales, and recent methods such as METOR Wang et al. (2025) and HyperGLM Nguyen et al. (2025) further incorporate richer object–relation dependencies and higher-order relational structures. Related works also study referring relationships Krishna et al. (2018); Kumar & Mishra (2023) and pixel-level video scene graphs Yang et al. (2023), where specified relations are used to localize interacting entities or their trajectories. Despite these advances, existing interaction modeling methods mainly target relation recognition or scene-graph prediction. In InterRVOS Jin et al. (2026), however, the interaction must further determine the asymmetric Actor and Target roles for segmentation. Existing methods do not explicitly study how relational information should be used to distinguish the two interacting participants by their roles. We make the first attempt to explicitly model how interaction information guides Actor–Target role reasoning in InterRVOS, through directed role propagation and triadic cross-role alignment.

3 Methodology

3.1 Overview

Given an input video and a natural-language expression describing an interaction, InterRVOS requires separate segmentation of the Actor and Target according to their asymmetric roles. The model predicts role-specific temporal masks:

$$\mathcal{Y} = f_\theta(V, E) = \left(\left\{ \widehat{M}_i^A \right\}_{i=1}^{N_A}, \left\{ \widehat{M}_j^T \right\}_{j=1}^{N_T} \right), \tag{1}$$

where $V = \{I_t\}_{t=1}^{T}$ denotes the input video, E denotes the interaction expression, and $\widehat{M}_i^A, \widehat{M}_j^T \in \{0,1\}^{T\times H\times W}$ represent the predicted temporal masks of the Actor and Target instances, respectively. N_A and N_T denote the numbers of predicted Actor and Target instances.

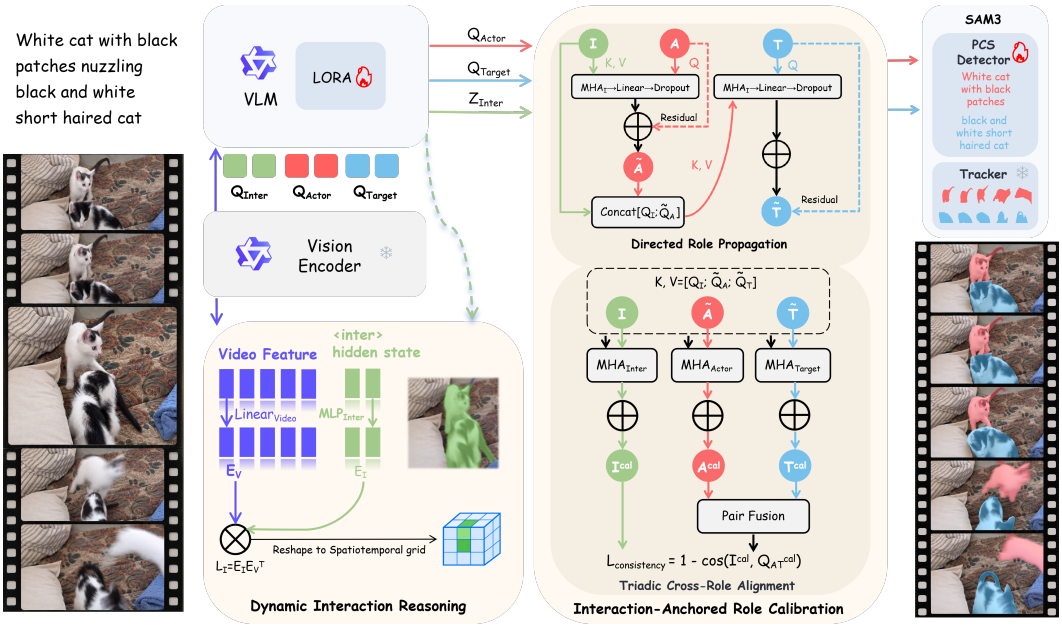

Figure 2: **Overview of our proposed R^{3}Refer.** R^{3}Refer consists of three main stages: (a) Dynamic Interaction Reasoning (DIR), which extracts Interaction, Actor, and Target queries from Qwen3-VL and grounds the interaction queries with video visual features to form a latent interaction representation; (b) Interaction-Anchored Role Calibration (IARC), which calibrates the role queries through Directed Role Propagation (DRP) and Triadic Cross-Role Alignment (TCRA); and (c) query-conditioned SAM3 Promptable Concept Segmentation (PCS) decoding and tracking, which uses the calibrated Actor and Target queries to produce role-specific video masks and tracks.

R^{3}Refer realizes this formulation through relation-conditioned role binding. It first constructs a visually grounded latent interaction state with *Dynamic Interaction Reasoning*. *Interaction-Anchored Role Calibration* then uses this state to establish the Actor representation and condition the complementary Target representation, followed by joint refinement of the Interaction, Actor, and Target queries. The calibrated role queries are finally decoded by the SAM3 PCS detector into role-specific temporal masks.

3.2 Dynamic Interaction Reasoning

Because Actor and Target roles are defined by their interaction rather than by object identity alone, role binding requires a relation-level state that captures how the described interaction is instantiated in the current video. To this end, R^{3}Refer introduces a group of `<inter>` queries into the multimodal sequence of Qwen3-VL. Together with the Actor and Target queries, they participate in multimodal reasoning over the video and interaction expression. Let $\mathbf{Q}_I$, $\mathbf{Q}_A$, and $\mathbf{Q}_T$ denote the Interaction, Actor, and Target query groups, respectively. The interaction-query hidden states are extracted as

$$\mathbf{H} = \mathcal{F}_{\text{Qwen3-VL}}\left(V, E; \mathbf{Q}_I, \mathbf{Q}_A, \mathbf{Q}_T\right), \qquad \mathbf{Z}_I = \mathbf{H}[\mathcal{P}_I], \tag{2}$$

where $\mathcal{F}_{\text{Qwen3-VL}}$ denotes the multimodal vision-language model and $\mathcal{P}_I$ represents the positions of the `<inter>` queries.

The resulting $\mathbf{Z}_I$ serves as an input-dependent latent relational state rather than a fixed action embedding, preserving the visual configuration under which the Actor and Target roles should be interpreted. To visually ground this state, R^{3}Refer projects the interaction hidden states and Qwen3-VL video features into a shared mask-embedding space:

$$\mathbf{E}_I = \text{MLP}_{\text{inter}}(\mathbf{Z}_I), \qquad \mathbf{E}_V = \text{Linear}_{\text{video}}\left(\text{VisionEncoder}_{\text{Qwen3-VL}}(V)\right). \tag{3}$$

The interaction mask logits are computed as $\mathbf{L}_I = \mathbf{E}_I \mathbf{E}_V^\top$. The logits are reshaped according to the spatiotemporal video-token grid and supervised using interaction regions. This supervision anchors

the latent interaction state to the participating visual evidence, providing the relational context for subsequent role binding.

3.3 Interaction-Anchored Role Calibration

Given the grounded interaction state, *Interaction-Anchored Role Calibration* performs relation-conditioned role binding in two steps. Directed Role Propagation first resolves the asymmetric Actor–Target dependency, after which Triadic Cross-Role Alignment jointly reconciles the interaction and its two role representations.

Directed Role Propagation. We first perform a directed query cascade that follows the asymmetric structure of the interaction. The Actor queries attend to the interaction queries to establish an Actor representation under the specified relation. The Target queries then attend to both the interaction queries and the updated Actor queries, using the established Actor context to constrain the complementary Target role. The two updates are defined as

$$\begin{aligned} \widetilde{\mathbf{Q}}_A &= \mathbf{Q}_A + \mathrm{MHA}_A\left(\mathbf{Q}_A, \mathbf{Q}_I, \mathbf{Q}_I\right), \\ \widetilde{\mathbf{Q}}_T &= \mathbf{Q}_T + \mathrm{MHA}_T\left(\mathbf{Q}_T, [\mathbf{Q}_I; \widetilde{\mathbf{Q}}_A], [\mathbf{Q}_I; \widetilde{\mathbf{Q}}_A]\right), \end{aligned} \tag{4}$$

where the interaction and updated Actor queries jointly serve as keys and values for Target reasoning. This directed propagation binds the two roles under a shared interaction rather than treating them as independent segmentation targets. Standard projection, normalization, and dropout operations are omitted from the formulation for clarity.

Triadic Cross-Role Alignment. After directed role propagation, the interaction and participant representations remain mutually informative. We therefore concatenate the interaction queries and calibrated role queries into a joint context, $\mathbf{Q}_{\mathrm{joint}} = [\mathbf{Q}_I; \widetilde{\mathbf{Q}}_A; \widetilde{\mathbf{Q}}_T]$. Each query group then reads from this shared context:

$$\mathbf{Q}_g^{+} = \mathbf{Q}_g + \mathrm{MHA}_g\left(\mathbf{Q}_g, \mathbf{Q}_{\mathrm{joint}}, \mathbf{Q}_{\mathrm{joint}}\right), \qquad g \in \{I, A, T\}. \tag{5}$$

This reciprocal refinement allows the grounded interaction and its bound roles to exchange complementary evidence and reconcile their representations before mask decoding.

3.4 Query-conditioned Video Mask Decoding

The calibrated Actor and Target queries are projected into the query feature space of the SAM3 PCS detector, while the corresponding role phrases are projected into its text-prompt space as auxiliary semantic conditions. PCS independently decodes each role query into a candidate mask and instance score, allowing multiple Actor or Target instances associated with the same interaction to be predicted in parallel. Actor and Target are decoded separately with their respective query groups and role phrases.

Compared with the prompt-driven SAM3 PVS mode, PCS assigns an independent query slot to each potential instance, making it more suitable for InterRVOS scenarios containing multiple interacting participants and reducing missed instances caused by a single spatial prompt.

3.5 Training and Inference

Training. R^{3}Refer jointly optimizes language modeling, interaction prediction, role-specific segmentation, and interaction–role consistency. For each branch $r \in \{I, A, T\}$, predictions are matched to the corresponding annotations using Hungarian matching; the interaction branch uses Actor–Target union tracks, while the Actor and Target branches use their respective role annotations. Each branch is optimized with classification, mask BCE, and Dice losses:

$$\mathcal{L}_r = \mathcal{L}_{\mathrm{cls}}^{r} + \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}}^{r} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}}^{r}, \qquad r \in \{I, A, T\}. \tag{6}$$

Here, $\mathcal{L}_{\mathrm{cls}}^{I}$ supervises valid interaction tracks, while $\mathcal{L}_{\mathrm{cls}}^{A}$ and $\mathcal{L}_{\mathrm{cls}}^{T}$ supervise Actor and Target query validity.

To align the interaction state with its bound roles, we fuse matched Actor and Target queries into an Actor–Target pair representation $\mathbf{Q}_{AT}^{\text{pair}}$ and impose a correspondence-aware consistency objective:

$$\mathcal{L}_{\text{con}} = 1 - \cos\left(\mathbf{I}^{\text{cal}}, \mathbf{Q}_{AT}^{\text{cal}}\right). \tag{7}$$

where $\mathbf{Q}_{I}^{\text{cal}}$ denotes the calibrated interaction queries, $\text{sim}(\cdot,\cdot)$ measures their correspondence with the Actor–Target pair representations, and $\mathbf{S}_{\text{gt}}$ is the binary ground-truth correspondence matrix.

The complete training objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_I\mathcal{L}_I + \lambda_A\mathcal{L}_A + \lambda_T\mathcal{L}_T + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{8}$$

where $\mathcal{L}_{\text{LM}}$ denotes the autoregressive language modeling objective, $\mathcal{L}_I$ is the interaction objective, and $\mathcal{L}_A$ and $\mathcal{L}_T$ are the Actor and Target segmentation objectives. The Interaction, Actor, and Target query tokens are excluded from next-token supervision.

Inference. During inference, the calibrated Actor and Target queries and their corresponding role phrases initialize the complete SAM3 video model. The PCS detector produces role-specific detections, which are associated and propagated through the video by the original SAM3 tracker. We retain detections above $\tau_{\text{det}} = 0.3$ and use $\tau_{\text{new}} = 0.5$ for newly detected objects.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We evaluate R^{3}Refer primarily on InterRVOS-127K (Jin et al., 2026), which contains 8,738 videos and 127,236 expressions. It provides 8,000 training videos with 122,188 expressions and 738 evaluation videos with 5,048 expressions across four settings: single object, multiple object, bidirectional interaction, and unidirectional interaction. Only the unidirectional setting includes separate Actor and Target annotations, while the others follow the Actor only RVOS protocol. We report region similarity J, contour accuracy F, and their mean $J\&F$ (Perazzi et al., 2016). Actor and Target results are reported separately for unidirectional expressions. For cross dataset generalization, we evaluate on Ref-DAVIS (Khoreva et al., 2018) and Ref-YouTube-VOS (Seo et al., 2020) for simple single object expressions, and on MeViS (Ding et al., 2023) and ReVOS (Yan et al., 2024a) for complex expressions involving multiple objects.

Implementation Details. We use Qwen3-VL-2B-Instruct (Bai et al., 2025a) as the vision language backbone and SAM3 (Carion et al., 2026) as the segmentation decoder. R^{3}Refer uses ten interaction queries and ten segmentation queries for each of the Actor and Target roles. The input expression also conditions role phrase generation, whose Actor and Target phrases are projected into the SAM3 PCS text prompt space. We fine tune R^{3}Refer with LoRA (Hu et al., 2021) using rank 64, scaling factor 16, and dropout 0.05. The training objective combines language modeling, interaction mask prediction, and role specific segmentation terms with coefficients $\lambda_{\text{LM}} = 1$, $\lambda_I = 1$, and $\lambda_A = \lambda_T = 0.5$. Within each prediction branch, the classification, binary cross entropy, and Dice terms are weighted by $\lambda_{\text{cls}} = 1$, $\lambda_{\text{bce}} = 2$, and $\lambda_{\text{dice}} = 0.5$, respectively.

4.2 Experimental Results

Quantitative Comparison. Table 1 compares R^{3}Refer with representative traditional RVOS methods (Wu et al., 2022; Ding et al., 2023), LLM based RVOS methods (Yan et al., 2024b; Bai et al., 2024; Yuan et al., 2025; Liu et al., 2026; Zhang et al., 2026; Gong et al., 2025; Hong & Do, 2026) and InterRVOS methods (Jin et al., 2026) developed for the InterRVOS task under the Actor, Target and standard RVOS settings. R^{3}Refer with Qwen3-VL-8B (Bai et al., 2025a) achieves the best results across all three settings and all evaluation metrics. It improves over VIRST (Hong & Do, 2026) by 3.6% on Actor and 5.2% on standard RVOS, and over ReVIOSa 4B (Jin et al., 2026) by 7.3% on Target. Even with Qwen3-VL-2B (Bai et al., 2025a) and Qwen2.5-VL-7B (Bai et al., 2025b), R^{3}Refer outperforms the other competing methods. These results demonstrate the benefit of its query based segmentation interface and the complementary roles of Dynamic Interaction Reasoning (DIR) and Interaction Anchored Role Calibration (IARC), which provide relation aware conditions for SAM3 (Carion et al., 2026).

Table 1: Performance comparison with existing methods on InterRVOS-127K.

Method	Backbone	InterRVOS-Actor			InterRVOS-Target			RVOS		
		J	F	$J\&F$	J	F	$J\&F$	J	F	$J\&F$
ReferFormer[CVPR'22]	Video-Swin-Base	59.1	59.9	59.5	10.3	12.6	11.5	52.0	53.2	52.6
LMPM[ICCV'23]	Tiny Swin Transformer	51.1	54.1	52.6	17.9	20.4	19.2	45.1	48.3	46.7
VISA[ECCV'24]	Chat-UniVi-7B	57.8	57.6	57.7	14.2	24.2	19.2	49.2	50.4	49.8
VideoLISA[NeurIPS'24]	LLaVA-Phi-3-V-3.8B	68.4	68.0	68.2	8.5	17.7	13.1	61.5	61.9	61.7
Sa2VA[CVPR'25]	InternVL2-4B	69.6	72.3	71.0	21.2	27.2	24.2	58.1	61.0	59.6
UniPixel[NeurIPS'25]	Qwen2.5-VL-7B	69.2	70.9	70.1	37.1	42.8	40.0	61.7	64.5	63.1
StreamingRVOS[CVPR'26]	InternVL2.5-4B	68.5	70.4	69.4	21.8	28.2	25.0	58.4	61.6	60.0
Veason-R1[CVPR'26]	Qwen2.5-VL-7B	69.2	72.0	70.6	31.8	40.5	36.2	60.1	64.2	62.1
VIRST[CVPR'26]	VideoChat-Flash-7B	74.4	76.3	75.3	25.7	31.0	28.4	65.1	68.0	66.5
ReVIOSa[CVPR'26]	InternVL-2.5-1B	71.8	74.7	73.3	65.9	68.9	67.4	60.2	63.8	62.0
ReVIOSa[CVPR'26]	InternVL-2.5-4B	73.2	75.8	74.5	67.1	69.5	68.3	63.0	66.1	64.5
R^{3}Refer(Ours)	Qwen2.5-VL-7B	76.3	77.9	77.1	72.4	75.7	74.1	68.1	70.8	69.5
R^{3}Refer(Ours)	Qwen3-VL-2B	76.9	78.4	77.7	72.9	76.0	74.5	68.5	71.3	69.9
R^{3}Refer(Ours)	Qwen3-VL-8B	**78.0**	**79.8**	**78.9**	**73.6**	**77.5**	**75.6**	**70.1**	**73.4**	**71.7**



Figure 3: Qualitative comparison of referred object segmentation on the RVOS subset of the InterRVOS-127K benchmark.

Qualitative Comparison. Fig. 3 illustrates qualitative comparisons between various models. Compared with representative recent RVOS methods (Liu et al., 2026; Gong et al., 2025; Hong & Do, 2026) and ReVIOSa (Jin et al., 2026), R^{3}Refer achieves more accurate interaction role segmentation and multi instance localization. This improvement mainly comes from Dynamic Interaction Reasoning and Interaction Anchored Role Calibration, which construct a relation level latent representation and transfer and align interaction information with the Actor and Target queries, providing more accurate role conditions for the SAM3 PCS detector.

Fig. 4 presents qualitative results from the unidirectional subset. Under the interaction expression "*Cat pushing toy*", ReVIOSa mislocalizes the toy, whereas R^{3}Refer correctly segments the Actor and Target. For the expression "*Person carrying camera and handbag*", which contains multiple Actor and Target pairs, ReVIOSa misses camera and handbag while R^{3}Refer successfully identifies the corresponding objects. These results highlight the effectiveness of Dynamic Interaction Rea-

Method	Actor	Target	RVOS
w/o DIR	69.8	65.2	63.5
Action Classification	70.3	65.8	63.9
Region Feature	75.2	69.9	68.3
Ours	78.9	75.6	71.7

Table 3: Ablation of dynamic interaction reasoning.

Method	DRP	TCRA	Actor	Target	RVOS
(i)	×	×	74.8	70.2	65.7
(ii)	×	✓	75.6	71.5	66.9
(iii)	✓	×	76.1	72.3	67.2
Ours	✓	✓	78.9	75.6	71.7

Table 4: Ablation of interaction-anchored role calibration.



Figure 4: Qualitative comparison of referred Actor and Target segmentation on the unidirectional subset of the InterRVOS-127K benchmark.

soning and Interaction Anchored Role Calibration in understanding interaction directionality and coordinating role representations.

4.3 Ablation Study

Analysis of Query Number. We investigate how the number of queries affects segmentation accuracy and inference efficiency by setting $K \in \{5, 10, 20\}$. As shown in Table 2, using $K = 5$ achieves the highest FPS but leads to noticeably lower performance on the Actor, Target and standard RVOS metrics. Increasing the number of queries to $K = 10$ yields the best segmentation results while maintaining a practical inference speed. Although $K = 20$ produces metrics comparable to those of $K = 10$, it substantially reduces the FPS because of the additional query computation. We therefore use $K = 10$ as the default setting in all subsequent experiments.

Table 2: Performance under different query numbers.

Query	FPS	Actor	Target	RVOS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
5	9.81	75.8	73.0	67.1
10	8.26	**78.9**	**75.6**	**71.7**
20	8.06	78.7	75.6	70.6

Effectiveness of Dynamic Interaction Reasoning. Fig. 5 shows that the interaction queries respond most strongly to the participating objects and their shared interaction region, indicating that Dynamic Interaction Reasoning learns a visually grounded representation of the described interaction. Removing Dynamic Interaction Reasoning reduces performance on the Actor, Target and conventional RVOS metrics. This indicates that the module performs visual interaction reasoning in the latent query space by modeling role relationships and refining query representations, thereby improving Actor and Target localization.

We compare supervision schemes for Dynamic Interaction Reasoning in Table 3. Removing DIR reduces $\mathcal{J}\&\mathcal{F}$ by 9.1, 10.4, and 8.2 points on Actor, Target, and RVOS, respectively, with the largest drop on Target, which depends most on relational context. Region Feature Supervision provides semantic cues but weaker pixel localization, whereas Action Classification Supervision captures interaction semantics with limited mask guidance and performs only slightly better than removing DIR. Interaction mask supervision is therefore better aligned with the final segmentation objective.



Figure 5: Interaction Query Response Maps computed from projected query embeddings and Qwen3-VL video features.

Table 5: Comparison of PCS and PVS on multi-object subsets.

Method	SAM type	InterRVOS-Multi	MeVIS-Multi(val^u)	ReVOS-Multi
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VeasonR1	PVS	68.9	69.5	60.3
VIRST	PVS	67.4	79.4	64.1
ReVIOSa	PVS	69.3	57.7	40.3
R^{3}Refer(SAM2)	PVS	63.9	75.8	64.8
R^{3}Refer(SAM3)	PVS	64.3	76.1	65.0
Ours	PCS	**85.7**	**83.3**	**70.9**

Effectiveness of Interaction-anchored Role Calibration. As shown in Table 4, both Directed Role Propagation and Triadic Cross Role Alignment improve role specific segmentation. Starting from the baseline without either component, Directed Role Propagation improves the Actor, Target, and RVOS results by 1.3%, 2.1%, and 1.5%, respectively, while Triadic Cross Role Alignment yields gains of 0.8%, 1.3%, and 1.2%. Directed Role Propagation transfers directional interaction information from the Interaction Query to the Actor and Target queries, with a stronger effect on Target prediction. Triadic Cross Role Alignment further coordinates the three query groups. When both components are enabled, R^{3}Refer achieves 78.9%, 75.6%, and 71.7% on Actor, Target, and RVOS, respectively, showing their complementary benefits.

Comparison Analysis of PCS and PVS. We compare PCS and PVS under the same language model and input expression. PCS passes multiple query representations to the segmentation decoder, whereas PVS uses a special token as the segmentation prompt. As shown in Table 5, replacing SAM2 with SAM3 under PVS brings only marginal gains (63.9 to 64.3 on InterRVOS-Multi), indicating that a stronger decoder alone is not the main source of improvement. Switching to PCS instead raises $\mathcal{J}\&\mathcal{F}$ by 21.4, 7.2, and 5.9 points on InterRVOS-Multi, MeViS-Multi, and ReVOS-Multi, respectively. Since each potential instance receives its own query slot, PCS avoids collapsing multiple participants into a single prompt, which is particularly beneficial when an interaction involves several Actors or Targets. Overall, R3Refer with PCS consistently outperforms methods with PVS across all multi-object subsets.

5 Conclusion

We propose R^{3}Refer, a novel framework that reformulates interaction-aware referring video object segmentation as relation-conditioned role binding. Dynamic Interaction Reasoning learns a visually grounded interaction representation that provides explicit relational context for role assignment. Building on this representation, Interaction-Anchored Role Calibration performs directed Actor–Target reasoning and jointly refines the Interaction, Actor, and Target representations through triadic alignment before mask decoding. Extensive experiments on InterRVOS-127K demonstrate that R^{3}Refer outperforms existing methods in both Actor and Target segmentation, highlighting the effectiveness of explicit interaction modeling for role-aware video segmentation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with part of code implementation, debugging, and the adaptation of evaluation scripts for InterRVOS-127K. We also used generative AI tools for language editing, including organizing, rephrasing, and polishing portions of the manuscript. Generative AI tools were not used to generate experimental results, determine research conclusions, or make final scientific judgments. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was checked against the implemented methods and experimental results. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies interaction-aware referring video object segmentation (InterRVOS) using existing public benchmark datasets. It does not involve human-subject experiments or the collection of new personal data. As with other video segmentation systems, the proposed method may produce incorrect object masks or Actor–Target role assignments, and such predictions should therefore be interpreted with appropriate caution in downstream applications. We have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing the methodological and experimental details necessary to reproduce the reported results. The main paper describes the model architecture, training objectives, datasets, evaluation protocols, metrics, and implementation settings. Additional implementation and experimental details are provided in the supplementary material. We intend to release the code and trained configurations upon publication.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4985–4995, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16372–16382, 2021.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023.

Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5119–5128. IEEE, 2023.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5958–5966, 2018.

Sitong Gong, Lu Zhang, Yunzhi Zhuge, Xu Jia, Pingping Zhang, and Huchuan Lu. Reinforcing video reasoning segmentation to think before it segments. *arXiv preprint arXiv:2508.11538*, 2025.

Jihwan Hong and Jaeyoung Do. Virst: Video-instructed reasoning assistant for spatiotemporal segmentation. *arXiv preprint arXiv:2603.27060*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

Xinjie Jiang, Chenxi Zheng, Xuemiao Xu, Bangzhen Liu, Weiying Zheng, Huaidong Zhang, and Shengfeng He. Vrdone: One-stage video visual relation detection. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 1437–1446, 2024.

Woojeong Jin, Seongchan Kim, Jaeho Lee, and Seungryong Kim. Interrvos: Interaction-aware referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10367–10376, 2026.

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)*, pp. 787–798, 2014.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian conference on computer vision*, pp. 123–141. Springer, 2018.

Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6867–6876, 2018.

Yogesh Kumar and Anand Mishra. Few-shot referring relationships in videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2289–2298. IEEE, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10840–10849, 2020.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems*, 38:126078–126108, 2026.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22803–22813. IEEE, 2023.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European conference on computer vision*, pp. 792–807. Springer, 2016.

Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. Hyperglm: Hypergraph for video scene graph generation and anticipation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29150–29160. IEEE, 2025.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 84–93, 2019.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pp. 208–223. Springer, 2020.

Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1300–1308, 2017.

Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10424–10433, 2019.

Yongqi Wang, Xinxiao Wu, and Shuo Yang. Metor: A unified framework for mutual enhancement of objects and relationships in open-vocabulary video visual relationship detection. *arXiv preprint arXiv:2505.06663*, 2025.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20193–20203. IEEE, 2025.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4964–4974. IEEE, 2022.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024a.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024b.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. Panoptic video scene graph generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18675–18685. IEEE, 2023.

Shuo Yang, Yongqi Wang, Xiaofeng Ji, and Xinxiao Wu. Multi-modal prompting for open-vocabulary video visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6513–6521, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Wenkang Zhang, Kaicheng Yang, Xiang An, Qiang Li, Ziyong Feng, Wankou Yang, and Jiankang Deng. Towards streaming referring video segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24598–24607, 2026.

Sipeng Zheng, Shizhe Chen, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18814–18824. IEEE, 2022.

A APPENDIX

Generalization to Other RVOS Datasets. Tables 6 and 7 evaluate the generalization of R^{3}Refer beyond InterRVOS. R^{3}Refer achieves the best result on Ref YouTube VOS with a score of 74.4% and remains competitive on Ref DAVIS with 78.2%. On the more challenging ReVOS benchmark, R^{3}Refer obtains 68.6%, 62.8%, and 65.7% on the Referring, Reasoning, and Overall subsets, respectively. These results show that the interaction aware query representation learned by R^{3}Refer transfers beyond InterRVOS and remains effective for both direct referring expressions and more complex reasoning based expressions.

Computational Efficiency. Table 8 compares the computational cost of R^{3}Refer with existing methods. The Qwen3-VL 2B version has the smallest model size and memory usage, requiring only $6{,}360$ MB and $9{,}083$ MB, respectively, although its lower throughput reflects the cost of video level interaction reasoning. Scaling to Qwen3-VL 8B increases the throughput to 8.36 FPS, which is comparable to the 7B VIRST baseline at 8.43 FPS. These results show that R^{3}Refer can operate with a compact 2B backbone while also scaling to a larger backbone without a substantial throughput reduction.

Table 6: Comparison with existing methods on Ref-DAVIS and Ref-YouTube-VOS.

Method	Ref-DAVIS			Ref-YouTube-VOS		
	J	F	$J\&F$	J	F	$J\&F$
ReferFormer	58.1	64.1	61.1	61.3	64.6	62.9
VISA	66.3	72.5	69.4	59.8	63.2	61.5
VideoLISA	64.9	72.7	68.8	61.7	65.7	63.7
InstructSeg	67.3	74.9	71.1	65.4	69.5	67.5
UniPixel	72.7	80.1	76.4	69.5	72.4	71.0
StreamingRVOS	72.2	81.0	76.6	68.4	72.7	70.5
VIRST	**75.9**	**83.1**	**79.5**	72.2	**76.1**	74.2
R^{3}Refer	75.8	80.6	78.2	**72.8**	75.9	**74.4**

Table 7: Comparison with existing methods on ReVOS.

Method	Referring			Reasoning			Overall		
	J	F	$J\&F$	J	F	$J\&F$	J	F	$J\&F$
ReferFormer	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1
LMPM	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4
VISA	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9
InstructSeg	54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5
UniPixel	63.9	67.8	65.8	59.4	63.7	61.5	61.7	65.7	63.7
StreamingRVOS	60.5	65.5	63.0	–	–	–	–	–	–
VeasonR1	60.3	65.6	63.0	53.6	60.0	56.8	56.9	62.8	59.9
VIRST	**68.8**	72.8	**70.8**	**63.9**	**68.3**	**66.1**	**66.3**	**70.6**	**68.4**
R^{3}Refer	64.0	**73.2**	68.6	61.2	64.3	62.8	62.6	68.8	65.7

Table 8: Comparison of computational efficiency and model size.

Method	Backbone	FPS	Memory (MB)	Model Size (MB)
Unipixel	7B	48.37	34,814	16,096
StreamingRVOS	4B	8.92	15,068	15,159
VeasonR1	3B	4.86	44,732	9,029
VIRST	7B	8.43	35,775	16,729
R^{3}Refer	2B	3.43	9,083	6360
R^{3}Refer	8B	8.36	39,464	18,157