InterRVOS: Interaction-Aware Referring Video Object Segmentation

Woojeong Jin Seongchan Kim Jaeho Lee Seungryong Kim

KAIST AI

Abstract

Referring video object segmentation (RVOS) aims to segment objects in a video described by a natural language expression. However, most existing approaches focus only on the referred object (typically the actor), even when the expression clearly describes an interaction involving multiple objects with distinct roles. In this paper, we introduce Interaction-Aware Referring Video Object Segmentation (InterRVOS), a novel task that focuses on explicit interaction modeling by requiring separate segmentation of actor and target objects. This formulation enables fine-grained understanding of object relationships, as many video events are defined by such interactions rather than individual objects. We present InterRVOS-127K, a large-scale dataset of over 127K automatically annotated expressions with distinct actor-target mask pairs, and propose ReVIOSa, a MLLM-based architecture that introduces interaction-aware special tokens and attention mask loss (AML) to enhance interaction-aware segmentation. We also propose a new evaluation protocol that separately evaluates actor and target segmentation for more accurate role distinction. Comprehensive experiments demonstrate that ReVIOSa outperforms existing baselines on the proposed InterRVOS-127K benchmark, with further analyses validating the necessity and effectiveness of both ReVIOSa and InterRVOS-127K. Our project page is available at: https://cvlab-kaist.github.io/InterRVOS.

1. Introduction

Referring video object segmentation (RVOS) aims to segment objects in a video that corresponds to a given referring expression. While earlier works for this task [5, 7, 14, 15, 23, 27, 28] primarily focused on aligning visual content with language to localize the referred object, recent advancements [6, 21, 29] have extended the scope of referring expressions to solve more challenging cases, such as motion-only cues or reasoning-based segmentation. These trends highlight a growing interest in capturing fine-grained temporal motions and achieving comprehensive video-language alignment.

Figure 1. **Task definition of InterRVOS.** We propose a novel task aiming to segment both actor and target objects separately from a given interaction expression, unlike standard RVOS approaches [2, 5, 6, 9, 15, 18, 27, 28, 30] that focus solely on the actor.

Despite these advances, one important yet underexplored aspect of RVOS is the understanding of *interactions* between objects. Standard RVOS [2, 5, 6, 9, 15, 18, 27, 28, 30] focuses on segmenting a single object or group of objects exhibiting similar motions, even if expressions that describe interactions with explicit actor and target are given. Such *interaction expressions* include not only the referred objects (actor), but also other objects involved in the interaction (target). For instance, an expression such as "A *reaching a hand towards* B" implies distinct semantic roles and spatio-temporal relationships between objects, where A is the *actor*, and B is the *target*. However, most existing RVOS approaches segment only the actor (A), neglecting the target object (B) involved in the interaction. Understanding such inter-object dynamics and the ability to distinguish between actor and target roles are essential, as many events in videos are defined not only by the motion of the object itself, but also by its relational context between objects.

Datasets	Annotation	Types			Statistics				
		Single	Multiple	Actor-Target	Video	Object	Expression	Obj/Video	Actor-Target
A2D Sentence [7]	Manual	✓	✗	✗	3,782	4,825	6,656	1.28	-
J-HMDB Sentence [7]	Manual	✓	✗	✗	928	928	928	1.00	-
Ref-DAVIS [14]	Manual	✓	✗	✗	90	205	1,544	2.27	-
Ref-Youtube-VOS [23]	Manual	✓	✗	✗	3,978	7,451	15,009	1.86	-
MeViS [6]	Manual	✓	✓	✗	2,006	8,171	28,570	4.28	-
ReVOS [29]	Manual	✓	✓	✗	1,042	5,535	35,074	5.31	-
Ref-SAV [30]	Automatic	✓	✗	✗	37,311	72,509	72,509	1.94	-
InterRVOS-127K	Automatic	✓	✓	✓	8,738	35,247	127,236	4.03	17,604

Table 1. **Comparison of existing RVOS datasets and InterRVOS-127K dataset.** Unlike existing datasets, InterRVOS-127K additionally supports interaction expressions (denoted as Actor-Target), which annotates separate masks of actor and target objects within an interaction. InterRVOS-127K contains over 127K mask-text pairs, and is the first to explicitly annotate the masks of both actors and targets.

In this work, we propose **Interaction-Aware Referring Video Object Segmentation (InterRVOS)**, a novel task that focuses on complex interaction scenarios within videos, requiring the model to segment the actor and target objects *separately*, as illustrated in Fig. 1. Importantly, this task explicitly models role directionality within interactions, capturing the asymmetry between actor and target. This formulation goes beyond segmenting all involved objects as a whole (i.e., union) by requiring the model to separately model each object's temporal behavior and to capture the inter-object dynamics that arise from their distinct roles. To enable evaluation under the InterRVOS setting, we introduce a new protocol that assesses segmentation performance for actor and target separately, for each interaction expression.

To support this task, we present **InterRVOS-127K**, a large-scale dataset containing over 127K expressions, automatically annotated using our data annotation pipeline. Unlike previous RVOS datasets [6, 7, 14, 23, 29], InterRVOS-127K includes separate mask annotations for actor and target objects for each interaction expression, enabling models to learn inter-object dynamics effectively. An overall comparison of datasets is provided in Tab. 1.

We further propose **ReVIOSa**, a novel multimodal large language model (MLLM)-based architecture designed to solve InterRVOS task. ReVIOSa introduces interaction-aware special tokens, `[SEG_ACT]` and `[SEG_TAR]`, which are passed to the mask decoder [22], as role-specific prompts to generate separate segmentation masks for actor and target objects, respectively. Unlike conventional approaches [2, 29–31] using only `[SEG]` token, which is limited to actor objects, our interaction-aware special tokens enable the model to explicitly represent and distinguish the asymmetric roles inherent in interactions, moving beyond single-role (actor-only) segmentation to role-aware segmentation. To further ensure that these tokens effectively capture object information, we propose attention mask loss (AML), which supervises the attention maps of `[SEG_ACT]` and `[SEG_TAR]` toward their corresponding object regions. By encouraging such attention localization, AML helps the model learn more discriminative representations for each role, leading to improved understanding of inter-object dynamics and more accurate distinction of individual motion patterns within interactions.

To summarize, our main contributions are as follows:

- We introduce a new task, InterRVOS, which focuses on interaction scenarios within videos by requiring distinct segmentation masks of both actor and target objects. We also propose an evaluation protocol that assesses segmentation performance separately for actor and target from an interaction expression.
- We present InterRVOS-127K, a large-scale dataset, containing 127K expressions including interaction expressions with distinct annotations of actor-target.
- We propose ReVIOSa, a novel MLLM-based architecture that incorporates interaction-aware special tokens and employs attention mask loss to improve role-aware segmentation required in InterRVOS.
- ReVIOSa achieves state-of-the-art results on InterRVOS-127K, demonstrating its effectiveness in modeling interactions and precise understanding of complex temporal motions.

2. Related work

Referring video object segmentation. RVOS aims to segment a referred object in a video given a natural language expression. Early works [3, 5, 7, 15, 19, 27, 28] mainly focused on appearance-based reasoning through multimodal fusion, often in single-frame or single-object settings. The introduction of MeViS [6] emphasized the importance of motion-aware and spatio-temporal reasoning by including motion-only and multi-instance expressions, prompting models to better track objects over time. Recent approaches [9, 18] adopt lightweight text-encoder-based frameworks, while others [2, 30, 31] leverage multimodal large language models (MLLMs) [16] and use spe-

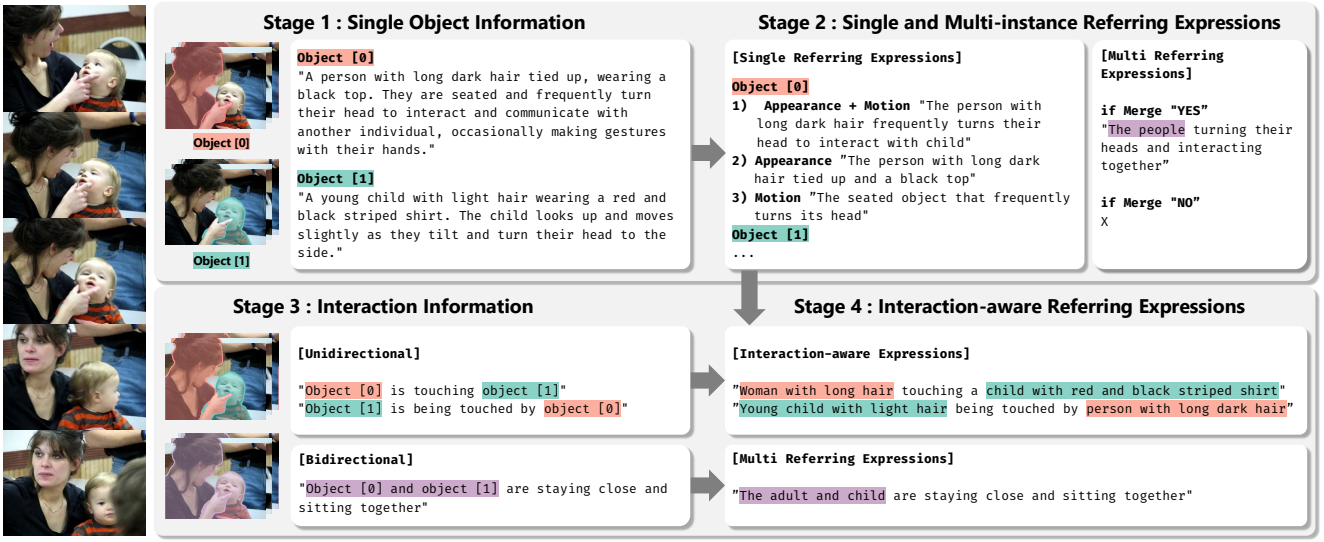


Figure 2. **Data annotation pipeline of InterRVOS-127K dataset.** Our proposed automatic data annotation pipeline constructs referring expressions for single, multi-object, and interaction scenarios in four stages. These stages sequentially extracts object appearance and motion, detects interactions, and generates detailed expressions grounded in both visual properties and interaction context.

cial tokens (e.g., `[SEG]`) to guide segmentation.

Despite these advances, existing methods remain actor-focused, performing segmentation solely on a single object (or group of objects) even when interaction expressions involve distinct actor and target roles inherently. While tasks like ReasonVOS [29] and Grounded Conversation Generation (GCG) [20, 21] move beyond traditional RVOS, they fall short in modeling interactions with directions between multiple objects. In particular, GCG treats segmentation as a noun phrase grounding problem without capturing interaction semantics such as role asymmetry.

In contrast, InterRVOS explicitly models the asymmetric roles within interactions by separating actor and target, demanding more precise and role-aware segmentation under interaction-aware expressions.

Video object interaction. Modeling object interactions in video requires a role-aware perspective that distinguishes actors from targets, as the semantics of relational events (e.g., *"person pushing cart"*) depend on how one object acts upon another. To support such modeling, prior works have introduced several datasets [13, 24, 25] with structured annotations, which are actor–predicate–target triplets over time, enabling models to capture visual relationships in dynamic contexts. More recent datasets like STAR [26] and MOMA [17] further incorporate temporal grounding and causal structure, capturing complex interactions.

These efforts highlight the importance of explicitly modeling inter-object dynamics as a foundation for fine-grained video understanding. However, existing datasets such as VidOR [25] or ActionGenome [13], which have been widely used in video scene graph generation (VSGG), are constructed with a closed set of predefined predicates and templated relation descriptions. As a result, they fundamentally lack the linguistic diversity and compositional expressiveness of natural language, making them insufficient for training models that aim to comprehend open-ended referring expressions or to address diverse interaction scenarios as in InterRVOS. To address these limitations, we introduce InterRVOS and dataset InterRVOS-127K, which provide interaction-aware annotations and require distinct segmentation of actor and target objects described in natural language interactions.

3. InterRVOS-127K dataset

While InterRVOS requires separate segmentation of actor and target objects, existing datasets [6, 7, 14, 23, 29, 30] lack annotations for the *target* object. To address this, we introduce **InterRVOS-127K**, an automatically annotated large-scale dataset which contains both conventional actor-only RVOS cases and interaction expressions with distinct actor-target annotations for both actor and target objects. Built upon VidOR [25], InterRVOS-127K is constructed via a stage-wise automatic annotation pipeline that leverages GPT-4o [12] and LLaMA-70B [8] to generate and verify high-quality captions. Additional details on InterRVOS-127K are provided in Appendix C.

3.1. Data annotation pipeline

To generate high-quality expressions which capture the precise interaction between actors and targets, we design a stage-wise automatic annotation pipeline consisting of four main stages. The overall data annotation pipeline is illustrated in Fig. 2.

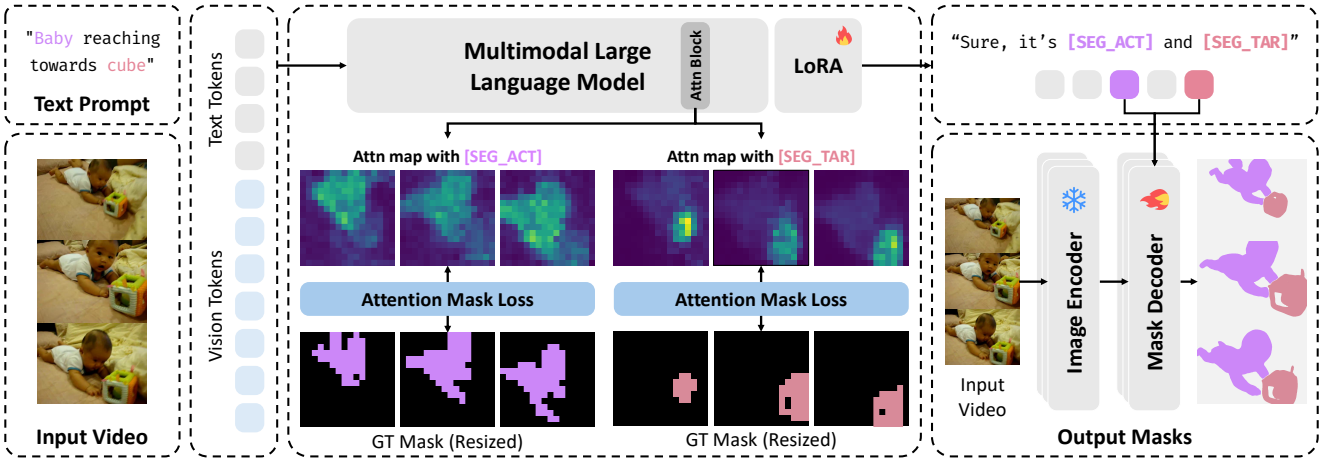


Figure 3. **Our proposed architecture.** ReVIOSa utilizes `[SEG_ACT]` and `[SEG_TAR]` tokens which explicitly separate the *actor* and the *target* within an interaction. Furthermore, it utilizes attention mask loss (AML) which supervises the attention maps of these tokens to align with their respective object regions. Together, these components enable ReVIOSa to effectively model interaction.

Prior to stage-wise processing, we pre-compute mask tracks for all objects in the video using SAM2 [22]. **Stage 1** captures each object's appearance and motion independently. **Stage 2** converts this into referring expressions, optionally merging descriptions for objects with similar motion patterns. **Stage 3** detects interactions, determines their role directionality, and assigns actor and target roles if unidirectional. **Stage 4** generates rich, interaction expressions by incorporating both class-level and appearance-specific cues, producing multiple paired expressions by swapping actor and target roles.

3.2. Training and evaluation set

Using data annotation pipeline, we automatically annotated 8,000 videos for training and 738 videos for evaluation. The numbers of expressions are 122,188 and 5,048, respectively. The evaluation set was manually verified and corrected by human annotators.

4. ReVIOSa architecture

We propose **ReVIOSa** (**Re**ferring **V**ideo **I**nteraction-aware **O**bject **S**egmenta**t**ion), a novel MLLM-based framework designed to effectively address the InterRVOS task. This task requires understanding asymmetric interactions, where the model must determine *which object acts upon which*, and segment actor and target objects separately based on their distinct roles in the interaction. To achieve this, ReVIOSa first introduces interaction-aware special tokens, `[SEG_ACT]` and `[SEG_TAR]`, which serve as explicit role representations within the MLLM. We further propose an attention mask loss (AML) that supervises the attention maps of these tokens to align with their respective object regions, ensuring that they effectively capture role-specific visual information. Together, these components enable precise modeling of inter-object dynamics and reliable role-aware segmentation. The overall architecture of ReVIOSa is shown in Fig. 3.

4.1. MLLM-based prompting

Given an input video $V = \{I_i\}_{i=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$ consisting of T frames and a referring expression E, our model aims to predict binary segmentation mask sequence $\hat{\mathcal{M}} = \{\hat{\mathcal{M}}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W}$, where each mask $\hat{\mathcal{M}}_t \in \{0,1\}^{H\times W}$ corresponds to the objects at time t. The overall framework consists of a LLaVA-based [16] multimodal large language model (MLLM) and a video segmentation model, SAM2 [22].

We first extract vision tokens $\mathbf{f}_{\text{vision}} \in \mathbb{R}^{N_v\times D_v}$ from a uniformly sampled video V' consisting of T' frames, and text tokens $\tilde{\mathbf{f}}_{\text{text}} \in \mathbb{R}^{N_t\times D}$ from the referring expression E, using the vision encoder and text tokenizer of the MLLM. Here, N_v and N_t denote the number of vision and text tokens, while D_v and D represent the embedding dimensions of the vision encoder and the MLLM, respectively. The vision tokens are projected into a shared embedding space with text tokens using MLP projection layer:

$$\tilde{\mathbf{f}}_{\text{vision}} = \text{MLP}_{\text{vision}}(\mathbf{f}_{\text{vision}}). \tag{1}$$

The projected vision tokens $\tilde{\mathbf{f}}_{\text{vision}} \in \mathbb{R}^{N_v\times D}$ and text tokens $\tilde{\mathbf{f}}_{\text{text}}$ are concatenated and fed into the MLLM $\mathcal{F}$ to produce the output sequence $\hat{\mathbf{y}}_{\text{out}}$:

$$\hat{\mathbf{y}}_{\text{out}} = \mathcal{F}([\tilde{\mathbf{f}}_{\text{vision}}; \tilde{\mathbf{f}}_{\text{text}}]), \tag{2}$$

where $\hat{\mathbf{y}}_{\text{out}}$ includes a special segmentation token, i.e., `[SEG]`. We extract the final-layer embedding $\tilde{\mathbf{h}}_{\text{seg}}$ corresponding to the `[SEG]` token and apply an MLP projection layer, MLP_{seg}, to obtain the prediction vector $\mathbf{p}_{\text{seg}} \in \mathbb{R}^{D_{\text{dec}}}$, where D_{dec} is the input embedding dimension of

the SAM2 mask decoder. In parallel, the vision encoder of SAM2 extracts visual features $\mathbf{v}_{\text{seg}} \in \mathbb{R}^{T \times N_{\text{enc}} \times N_{\text{enc}} \times D_{\text{enc}}}$ from the input video V, where $N_{\text{enc}} \times N_{\text{enc}}$ denotes the spatial resolution of the encoder feature map and D_{enc} is the corresponding feature dimension.

Finally, SAM2 mask decoder $\mathcal{F}_{\text{dec}}$ produces the binary mask sequence $\hat{\mathcal{M}}$. The overall process is formulated as:

$$\mathbf{p}_{\text{seg}} = \text{MLP}_{\text{seg}}(\tilde{\mathbf{h}}_{\text{seg}}), \quad \hat{\mathcal{M}} = \mathcal{F}_{\text{dec}}(\mathbf{v}_{\text{seg}}, \mathbf{p}_{\text{seg}}). \tag{3}$$

4.2. Interaction-aware special tokens

To enable role-aware reasoning and segmentation within interactions, we introduce interaction-aware special tokens, `[SEG_ACT]` and `[SEG_TAR]`, which explicitly represent the actor and target roles by segmenting their corresponding objects. Unlike standard `[SEG]` token approaches [2, 30, 31] that produce actor-only outputs, our role-specific tokens allow the model to represent the actor and target as distinct participants thus enhancing the model's ability to capture their relational context. Serving as explicit semantic anchors, these tokens provide separate role embeddings that help the model capture each object's characteristic motion and understand how their combined behaviors constitute the interaction.

Depending on the type of referring expression E, the model dynamically determines whether to generate one or two special tokens. At inference time, the model first determines whether the input expression involves an interaction. If so, it outputs both the `[SEG_ACT]` and `[SEG_TAR]` tokens for distinct segmentation. Otherwise, only the `[SEG_ACT]` token is generated for actor segmentation. In this new setting, the output of the MLLM $\hat{\mathbf{y}}_{\text{out}}$ can now include interaction-aware special tokens.

The corresponding hidden states for each special token $\tilde{\mathbf{h}}_{\text{act}}$ and $\tilde{\mathbf{h}}_{\text{tar}}$ at the last layer of the MLLM are projected into SAM2's prompt embedding space:

$$\mathbf{p}_{\text{act}} = \text{MLP}_{\text{seg}}(\tilde{\mathbf{h}}_{\text{act}}), \quad \mathbf{p}_{\text{tar}} = \text{MLP}_{\text{seg}}(\tilde{\mathbf{h}}_{\text{tar}}), \tag{4}$$

where $\mathbf{p}_{\text{tar}}$ is used only when `[SEG_TAR]` is generated. Finally, the segmentation mask outputs are computed as:

$$\hat{\mathcal{M}}_{\text{act}} = \mathcal{F}_{\text{dec}}(\mathbf{v}_{\text{seg}}, \mathbf{p}_{\text{act}}), \quad \hat{\mathcal{M}}_{\text{tar}} = \mathcal{F}_{\text{dec}}(\mathbf{v}_{\text{seg}}, \mathbf{p}_{\text{tar}}). \tag{5}$$

4.3. Attention mask loss

While the interaction-aware special tokens provide explicit role representations, it is crucial to ensure that these tokens effectively capture their corresponding object information for accurate interaction-aware segmentation. To achieve this, we propose an attention mask loss (AML) that supervises the visual attention maps of `[SEG_ACT]` and `[SEG_TAR]` tokens within MLLMs to align with their respective object regions. The brief concept of AML is illustrated in Fig. 4.

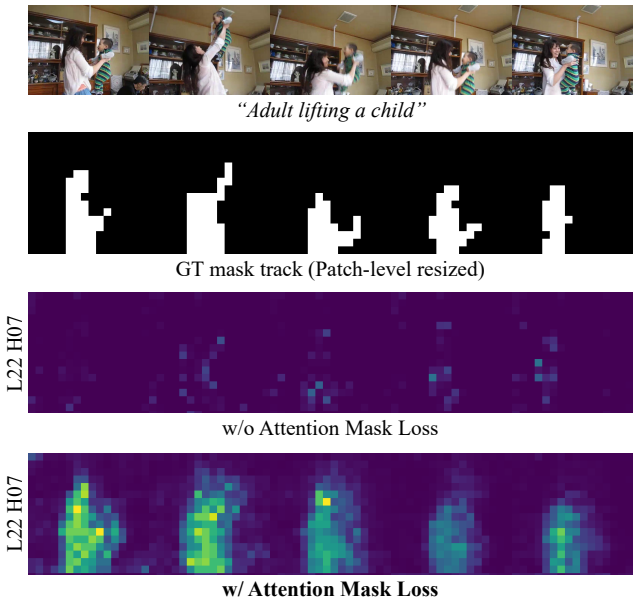


Figure 4. **Effectiveness of our proposed AML.** AML increases the proportion of attention with which the special token attends to the corresponding object region. L22H07 denotes the 7th head of the 22nd layer (indices start at 0).

During the generation of special tokens, the MLLM produces self-attention weight matrices at each transformer layer and head. Each attention map is of size $(N_v + N_t) \times (N_v + N_t)$, where $N_v = T' \times P \times P$ denotes the number of vision tokens and N_t is the number of text tokens. Here, $P \times P$ represents the number of patches per frame. From each attention map, we extract the vision attention weights from the special segmentation token (i.e., the query tokens `[SEG_ACT]` or `[SEG_TAR]`) to all vision tokens. These weights are then reshaped into a spatio-temporal attention map $A^{(l,h)} \in [0,1]^{T' \times P \times P}$ for each layer l and head h, aligning with the patch layout of the input video frames.

We first identify a set of specific layer-head pairs $\mathcal{H}$ using a selection protocol based on vision attention. For each selected $(l,h) \in \mathcal{H}$, we supervise the attention map $A^{(l,h)}$ using the ground-truth binary mask $\mathcal{M}' \in \{0,1\}^{T' \times H \times W}$, which is resized to the patch resolution, resulting in $\mathcal{G}' \in \{0,1\}^{T' \times P \times P}$. Since our method distinguishes between actor and target objects, we apply supervision to each type jointly. Specifically, the AML is defined as:

$$\mathcal{L}_{\text{AML}} = \sum_{r \in \{\text{act},\text{tar}\}} \sum_{(l,h) \in \mathcal{H}} \text{BCE}\left(A_r^{(l,h)}, \mathcal{G}'_r\right). \tag{6}$$

The summation over $r \in \{\text{act}, \text{tar}\}$ applies only to the roles present in the expression; if an expression contains only `[SEG_ACT]`, the term for `[SEG_TAR]` is dismissed.

By guiding each token's attention toward its corresponding object region, AML encourages `[SEG_ACT]` and `[SEG_TAR]` to learn accurate role-aware representations.

Methods	InterRVOS-Actor			InterRVOS-Target			RVOS		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Referformer [28]	59.1	59.9	59.5	-	-	-	52.0	53.2	52.6
LMPM [6]	51.1	54.1	52.6	-	-	-	45.1	48.3	46.7
VISA-7B [29]	57.8	57.6	57.7	-	-	-	49.2	50.4	49.8
VideoLISA-3.8B [2]	68.4	68.0	68.2	-	-	-	61.5	61.9	61.7
Sa2VA-1B [30]	69.9	72.6	71.3	-	-	-	55.4	58.7	57.0
Sa2VA-4B [30]	69.6	72.3	71.0	-	-	-	58.1	61.0	59.5
ReVIOSa-1B	71.8	74.7	73.3	65.9	68.9	67.4	60.2	63.8	62.0
ReVIOSa-4B	**73.2**	**75.8**	**74.5**	**67.1**	**69.5**	**68.3**	**63.0**	**66.1**	**64.5**

Table 2. **Quantitative results on InterRVOS-127K dataset.** ReVIOSa achieves the best performance on both interaction-aware (InterRVOS-Actor and InterRVOS-Target) and standard RVOS settings, highlighting its ability to effectively model complex motions. The best-performing results are presented in **bold**, while the second-best results are underlined.

Together with the interaction-aware tokens, AML strengthens their semantic roles to the correct visual regions, reinforcing role distinction and enabling precise modeling of inter-object dynamics. This auxiliary loss is optimized jointly with the segmentation loss during training.

4.4. Overall training loss

We apply standard pixel-wise cross-entropy loss and dice loss between the predicted mask $\hat{\mathcal{M}}$ and ground-truth mask track $\mathcal{M}$:

$$\mathcal{L}_{\text{seg}} = \sum_{r\in\{\text{act},\text{tar}\}} \mathcal{L}_{\text{CE}}(\hat{\mathcal{M}}_r, \mathcal{M}_r) + \mathcal{L}_{\text{Dice}}(\hat{\mathcal{M}}_r, \mathcal{M}_r). \tag{7}$$

When the referring expression describes an interaction, the segmentation loss is computed for both masks, $\hat{\mathcal{M}}_{\text{act}}$ and $\hat{\mathcal{M}}_{\text{tar}}$. Otherwise, it is computed only on $\hat{\mathcal{M}}_{\text{act}}$. We also include a text loss $\mathcal{L}_{\text{text}}$, defined as the cross-entropy loss over the predicted and ground-truth answer. Consequently, the total training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{AML}} \cdot \mathcal{L}_{\text{AML}} + \lambda_{\text{text}} \cdot \mathcal{L}_{\text{text}}, \tag{8}$$

where λ_{AML} and λ_{text} are weighting coefficients for the attention mask loss and text loss, respectively.

5. Experiments

In this section, we first outline the implementation details and training setup of our framework in Section 5.1. We then report quantitative, qualitative, and ablation results of ReVIOSa on the InterRVOS-127K benchmark in Section 5.2. Finally, we present analysis experiments that examine the necessity of the InterRVOS task and validate the effectiveness of both ReVIOSa and InterRVOS-127K, along with additional in-depth analyses on AML in Section 5.3. Further experimental details are provided in the Appendix, including extended analysis of the proposed AML (Appendix A) and quantitative as well as qualitative results on standard RVOS benchmarks (Appendix B).

All evaluations follow standard RVOS metrics [6, 14, 23], using the average of region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, denoted as $\mathcal{J}\&\mathcal{F}$.

5.1. Implementation details

For the proposed architecture ReVIOSa, we utilize InternVL-2.5 [4] as the base model for multimodal large language model (MLLM), applying LoRA [10] tuning exclusively. For the segmentation module, we adopt SAM2 [22] and fine-tune only its decoder while keeping the image encoder frozen. The model is trained for 10 epochs with a batch size of 128. We report results using two model scales: 1B and 4B which are trained on 4 NVIDIA RTX 3090 GPUs for 12 hours and 4 NVIDIA A6000 GPUs for 16 hours, respectively. For loss weighting, we set $\lambda_{\text{AML}} = 1.0$ and $\lambda_{\text{text}} = 1.0$ for all experiments.

5.2. Experimental results

Quantitative results. Tab. 2 presents the quantitative results under three evaluation settings: InterRVOS-Actor, InterRVOS-Target, and RVOS. The first two are newly introduced evaluation protocols for InterRVOS, designed to assess the role-specific segmentation of actor and target objects for each interaction expression. RVOS represents the standard setting which only segments the referred (actor) objects, which incorporates all expression samples. In these evaluation settings, the input prompt is formatted as *"Please segment {expression} in this video."* and ReVIOSa is trained to output interaction-aware special tokens when the referring expression includes interaction. For ReVIOSa, we use the `[SEG_ACT]` token for InterRVOS-Actor and RVOS, and the `[SEG_TAR]` token for InterRVOS-Target. All other baseline models are evaluated with a single conventional `[SEG]` token.

Importantly, previous RVOS approaches [6, 28–30] are designed to segment only the actor, and thus are not applicable to the InterRVOS-Target setting, highlighting the novelty and necessity of our proposed task. Even so, our

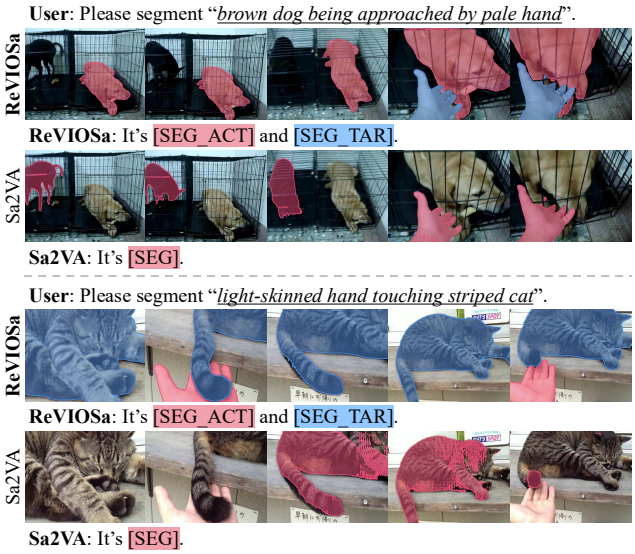


Figure 5. **Qualitative results.** Compared to the previous RVOS method, Sa2VA [30], ReVIOSa accurately segments both the actor and the target objects when given an interaction expression, demonstrating its ability to distinguish object roles.

	[SEG_ACT] [SEG_TAR]	AML	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
(i)	✗	✗	55.4	58.7	57.0
(ii)	✗	✓	57.4	59.6	58.5
(iii)	✓	✗	57.8	61.3	59.6
(iv)	✓	✓	**60.2**	**63.8**	**62.0**

Table 3. **Ablation study on ReVIOSa architecture.** Both (ii) AML and (iii) the interaction-aware special tokens contribute performance gains over (i) the baseline. (iv) ReVIOSa achieves the highest performance among all configurations.

proposed ReVIOSa demonstrates competitive performance on both the InterRVOS-Actor and RVOS. The 1B model already surpasses previous methods on most metrics, while the 4B model achieves state-of-the-art performance across all metrics. This indicates that role-specific supervision in our framework not only facilitates interaction modeling for segmenting the actor in interaction expressions but also improves performance on conventional RVOS, demonstrating its broad benefit in relational understanding and complex motion reasoning.

Qualitative results. Fig. 5 compares qualitative results under the InterRVOS setting, where both the input video and expression involve multiple interacting objects. In these complex cases, the previous state-of-the-art RVOS method, Sa2VA [30], fails to identify the referred object under interaction. In contrast, ReVIOSa is explicitly trained to distinguish object roles, enabling more precise recognition and segmentation of both actor and target objects.

Ablation studies. We perform an ablation study to assess the individual and combined contributions of two core components: the interaction-aware special tokens ([SEG_ACT] and [SEG_TAR]) and the proposed attention mask loss (AML). As presented in Tab. 3, each component independently improves model performance over the (i) baseline (57.0 $\mathcal{J}\&\mathcal{F}$), which uses conventional [SEG] and no AML. Specifically, AML contributes +1.5, while the interaction-aware tokens add +2.6. When combined, the model achieves a performance of 62.0 on the InterRVOS-127K evaluation set. This improvement demonstrates that semantic role separation through special tokens and direct attention supervision through AML complement each other effectively, resulting in better interaction modeling and the best overall performance.

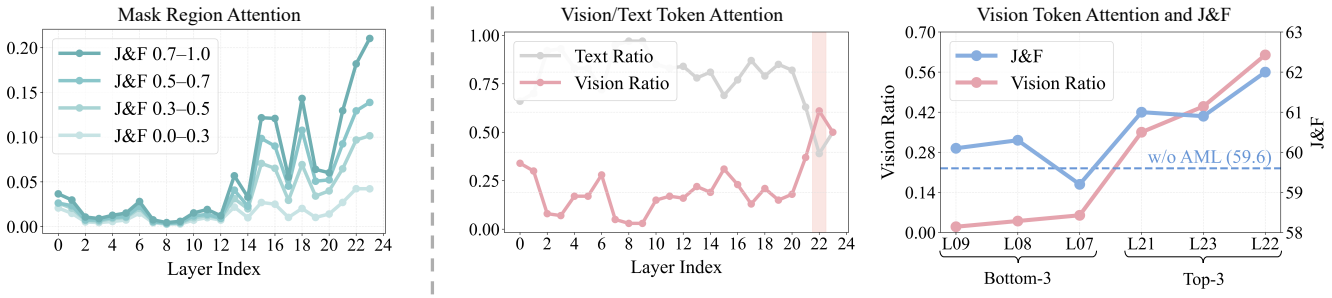
Methods	InterRVOS-Target		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
VISA-7B [29]	14.2	24.2	19.2
VideoLISA-3.8B [2]	8.5	17.7	13.1
Sa2VA-4B [30]	21.2	27.2	24.2

Table 4. **Zero-shot evaluation of existing RVOS models on InterRVOS-Target setting.** The results show that current RVOS methods fail to generalize to interaction-centric scenarios, highlighting the necessity of our InterRVOS task formulation, the InterRVOS-127K dataset, and the ReVIOSa architecture.

5.3. Analysis

Necessity of InterRVOS. To assess whether existing referring video object segmentation (RVOS) models can handle interaction expressions, we evaluate several MLLM-based baselines [2, 29, 30] using a modified prompt that instructs the models to reason about the target objects. Specifically, we use *"Please segment the object that is directly interacting with {expression} in the image."*. This requires the model to distinguish between the *actor* (the referred object) and the *target* (the object being interacted with), and to correctly infer the directionality of interaction. As shown in Tab. 4, despite MLLM's strong reasoning capability acquired from large-scale multimodal training [1, 11], all baseline models perform poorly under this setting, failing to identify the corresponding target object involved in the interaction and revealing their limited ability to capture the directionality and semantics of inter-object dynamics. These results clearly indicate the necessity of InterRVOS, which explicitly supervises interaction modeling.

Effectiveness of dataset and model. Tab. 5 presents zero-shot evaluation results of models trained on different datasets, ReVOS [29], Ref-SAV [30], and InterRVOS-127K, on three standard RVOS benchmarks: MeViS [6], Ref-Youtube-VOS [23], and Ref-DAVIS [14]. These results illustrate how much transferable video understanding each training dataset provides. The baseline matches the setting used in Tab. 3 (i). Notably, the baseline model trained on our large-scale and well-curated InterRVOS-127K, constructed



(a) Correlation between attention and J&F (b) Layer-wise Vision/Text Attentions (c) Layer-wise impact of AML

Figure 6. **Motivation and verification of AML.** (a) Segmentation accuracy increases when the segmentation token attends more strongly to the ground-truth mask region. (b) Layers exhibit varying degrees of vision attention. (c) Applying AML to layers with high vision attention yields clear gains, with Layer 22 achieving the largest improvement (+2.4 $\mathcal{J}\&\mathcal{F}$).

Dataset	Baseline			ReVIOSa
	ReVOS [29]	Ref-SAV [30]	InterRVOS-127K	InterRVOS-127K
MeViS valid [6]	39.6	32.8	40.4	**42.4**
MeViS valid_u [6]	49.1	40.1	49.5	**50.1**
Ref-Youtube-VOS [23]	57.5	54.2	**61.2**	60.3
Ref-DAVIS [14]	62.8	62.1	65.9	**66.2**

Table 5. **Zero-shot evaluation on standard RVOS benchmarks.** Baseline results show that training on our InterRVOS-127K dataset yields stronger transferability than ReVOS [29] or Ref-SAV [30]. Combining our dataset with the proposed ReVIOSa architecture further achieves the best performance across all benchmarks.

through a fully automatic generation pipeline, achieves the highest performance across all benchmarks, outperforming the same model trained on ReVOS or Ref-SAV. We also evaluate our proposed architecture, ReVIOSa-1B, trained on InterRVOS-127K. By explicitly modeling interaction, ReVIOSa-1B achieves the best performance across all benchmarks, surpassing the baseline model trained on the same dataset. This demonstrates that combining our interaction-centric dataset with our interaction-aware architecture yields the most effective transferability to standard RVOS benchmarks.

Motivation of AML. For the analysis of AML, all results in Fig. 6 are generated using the ReVIOSa-1B model on the InterRVOS-127K benchmark with `[SEG_ACT]` token (details in the Appendix A). We analyze how the token's attention patterns across different layers relate to performance and inform our design choices for AML. For this analysis, attention values at each layer are computed by averaging across all attention heads. Additional details and head-wise analyses for head selection are provided in Appendix A.

As shown in Fig. 6 (a), we observe a clear correlation between segmentation performance and the `[SEG_ACT]` token's attention to the ground-truth mask region (denoted as Mask-Region Attention). Samples in which the token attends more strongly to the corresponding object region consistently achieve higher $\mathcal{J}\&\mathcal{F}$ scores. This suggests that segmentation accuracy, which reflects the model's understanding toward the object, depends on how well the token attends to mask-relevant regions. Building on this observation, we hypothesize that guiding attention behavior may benefit interaction modeling when jointly used with `[SEG_ACT]` and `[SEG_TAR]` we proposed. Specifically, when each token attends to the object region, it can develop a more focused understanding of the object's visual features and motion patterns, which leads to better distinction between actor and target roles and more accurate modeling of their interaction.

Layer selection for AML. We investigate where to apply AML within the MLLM. We hypothesize that the effectiveness of AML depends on how much the token relies on visual information at each layer. To investigate this, we analyze the layer-wise attention to vision and text tokens (denoted as Vision/Text Token Attention). As shown in Fig. 6 (b), different layers exhibit varying degrees of vision attention. To validate the layer choice, we apply AML to top-3 and bottom-3 layers. As illustrated in Fig. 6 (c), the top-3 layers (L22, L23, L21) consistently outperform the bottom-3 layers (L09, L08, L07), confirming that AML is effective when applied to layers where the token attends more strongly to vision tokens. Among these, Layer 22 yields the highest gain of +2.4 in $\mathcal{J}\&\mathcal{F}$ over the baseline.

6. Conclusion

We present InterRVOS, a novel task that focuses on the complex interaction cases by requiring the segmentation of both actor and target objects from a single interaction expression, thereby explicitly modeling inter-object dynamics. To support this, we present InterRVOS-127K, a large-scale dataset with over 127K expressions and distinct actor-target annotations for interaction expressions. We also propose ReVIOSa, a novel MLLM-based model with interaction-aware tokens and attention mask loss for precise role-specific segmentation. Extensive experiments validate the effectiveness of modeling interaction, with ReVIOSa achieving state-of-the-art performance on InterRVOS-127K.

Acknowledgment This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS-2025-25441838, RS-2025-02214479, RS-2025-02217259) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068, RS-2023-00222280, RS-2023-00266509), and National Research Foundation of Korea (RS-2024-00346597).

References

[1] Abdelrahman Abdelhamed, Mahmoud Afifi, and Alec Go. What do you see? enhancing zero-shot image classification with multimodal large language models. *arXiv preprint arXiv:2405.15668*, 2024. 7

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 1, 2, 5, 6, 7

[3] 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*, pages 4985–4995, 2022. 2

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6

[5] 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

[6] 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, 6, 7, 8

[7] 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. 1, 2, 3

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[9] 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

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[11] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 7

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[13] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10236–10247, 2020. 3

[14] 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, 2, 3, 6, 7, 8

[15] 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, 2

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 4

[17] Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. Moma: Multi-object multi-actor activity parsing. *Advances in neural information processing systems*, 34:17939–17955, 2021. 3

[18] 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. *Advances in Neural Information Processing Systems*, 36:26425–26437, 2023. 1, 2

[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. 2

[20] 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 Computer Vision and Pattern Recognition Conference*, pages 19036–19046, 2025. 3

[21] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1, 3

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4, 6

[23] 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*, pages 208–223. Springer, 2020. 1, 2, 3, 6, 7, 8

[24] 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*, pages 1300–1308, 2017. 3

[25] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 3

[26] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 3

[27] 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, 2

[28] 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, 6

[29] 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*, pages 98–115. Springer, 2024. 1, 2, 3, 6, 7, 8

[30] 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. 1, 2, 3, 5, 6, 7, 8

[31] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. Villa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23667–23677, 2025. 2, 5