InterRVOS: Interaction-aware Referring Video Object Segmentation – Supplementary Materials –

Overview

In the appendix, we provide additional details and analyses that further support the results and findings presented in the main paper.

- Sec. A provides additional analyses on attention mask loss (AML), including detailed observations of its behavior across layers and heads.
  - Layer-head selection for AML (Sec. A.1)
  - Attention comparison between interaction-aware special tokens (Sec. A.2)
  - Attention visualization (Sec. A.3)
- Sec. B presents additional experimental results demonstrating the effectiveness of ReVIOSa and InterRVOS-127K.
  - Quantitative results on RVOS benchmarks (Sec. B.1)
  - Comparison of training datasets (Sec. B.2)
  - ReVIOSa with other MLLM (Sec. B.3)
  - Additional qualitative results (Sec. B.4).
- Sec. C provides further details about InterRVOS-127K
  - Distribution analysis (Sec. A5)
  - Data annotation pipeline (Sec. C.2)
  - Additional dataset samples (Sec. C.3)
  - Video clip extraction procedure (Sec. C.4)
  - Overall dataset statistics (Sec. C.5).
- Sec. D discusses failure cases and outlines potential future directions for improving RVOS and our approach.

A. Analysis on AML

In this section, we present our analysis of the attention maps from the MLLM and the detailed layer-head selection process for both the 1B and 4B models. Specifically, Sec. A.1 describes how we select appropriate layer-head pairs based on special token's attention to vision tokens, and outlines the detailed selection strategies across model scales (1B and 4B). Sec. A.2 then compares the attention patterns of the interaction-aware special tokens, while Sec. A.3 analyzes the attention maps across various configurations.

A.1. Layer-head selection for AML

The attention weights across layers and heads for both 1B and 4B models are visualized in Fig. A1, where (a) and (b) correspond to the 1B model, and (c) and (d) to the 4B model. As described in the main paper, our layer-head selection strategy for AML first identifies the layer with the highest head-averaged attention to vision tokens, referred to as the *vision ratio*. For the 1B model, this peak appears at Layer 22 (Fig. A.1 (a)). We then analyze the head-wise attention scores within this layer (Fig. A.1 (b)), which is the layer with top-1 vision ratio, and empirically find that applying AML to the top-4 heads yields the best performance, as reported in Tab. A1. For the 4B model, Fig. A1 (c) shows that Layer 33 exhibits the highest vision ratio. Within this layer, we further analyze the head-wise attention scores (Fig. A1 (d)) and select the top-4 heads, H13, H12, H14, and H06, for AML supervision.

By consistently applying AML to layers and heads with strong attention to vision tokens, we effectively deliver spatial supervision to the most responsive components of the MLLM.

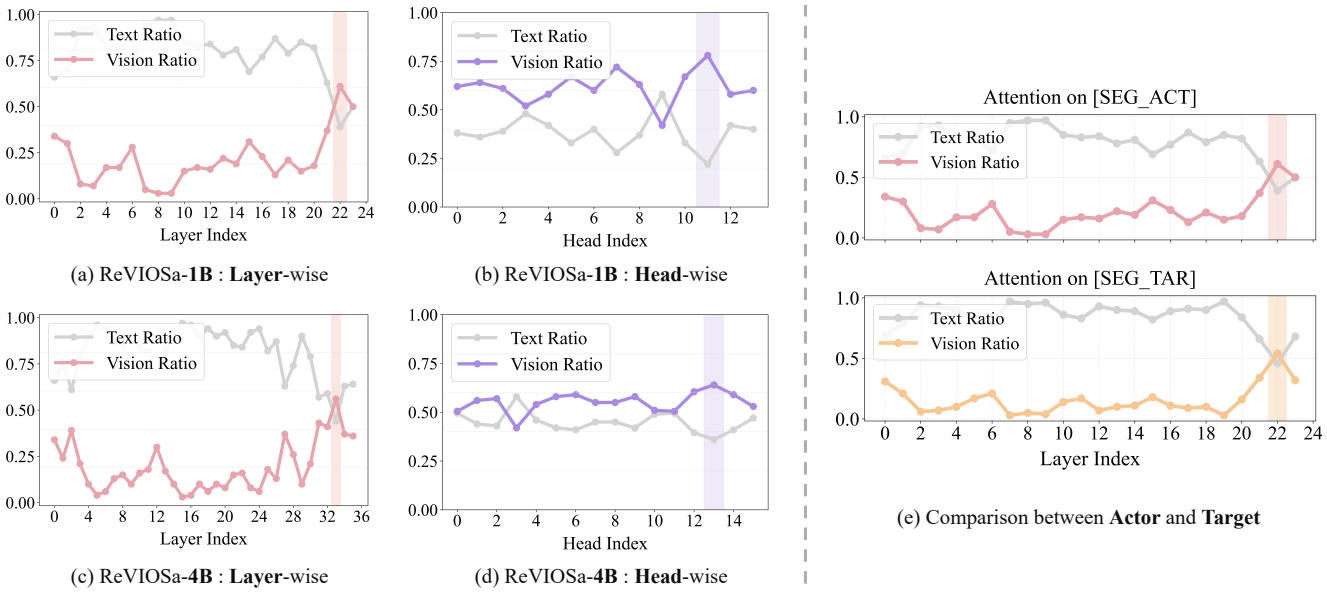
A.2. Attention comparison between interaction-aware special tokens

We compare the attention magnitude of the interaction-aware special tokens, `[SEG_ACT]` and `[SEG_TAR]`, to examine whether separate selection strategies are required for each token. In the main paper, all layer-head selections for AML were conducted based on the attention maps of the `[SEG_ACT]` token. This decision is justified by the observation that the attention patterns of both tokens are similar.

As shown in Fig. A1 (e), the top row presents the layer-wise attentions of the `[SEG_ACT]` token, while the bottom row shows the attentions for the `[SEG_TAR]` token. Notably, both tokens exhibit the highest vision attention at Layer 22, indicating that the same top-1 vision-attending layer is shared across the two roles. The distributions are closely aligned across layers, indicating that applying the selection strategy based on the `[SEG_ACT]` token is sufficient for effective supervision of both roles, without requiring additional role-specific tuning.

A.3. Attention visualization

Comparison of attention maps with and without AML. Fig. A2 illustrates the differences in attention maps between



(a) ReVIOSa-**1B** : **Layer**-wise (b) ReVIOSa-**1B** : **Head**-wise (c) ReVIOSa-**4B** : **Layer**-wise (d) ReVIOSa-**4B** : **Head**-wise (e) Comparison between **Actor** and **Target**

Figure A1. **Attentions across layers and heads.** Each figure illustrates the attention patterns from the special token (query) to vision tokens (key) across different layers and heads of the MLLM. (a) Layer-wise head-averaged attentions for the 1B model. (b) Head-wise attentions within Layer 22 of the 1B model. (c) Layer-wise head-averaged attentions for the 4B model. (d) Head-wise attentions within Layer 33 of the 4B model. (e) Layer-wise attention patterns of `[SEG_ACT]` (top) and `[SEG_TAR]` (bottom).

All Heads	$k=1$ (H11)	$k=2$ (+ H07)	$k=3$ (+ H10)	$k=4$ (+ H05)	$k=5$ (+ H08)
60.7	61.3	60.5	61.2	**62.0**	60.6

Table A1. **Performance comparison of AML applied to top-k attention heads in layer 22.** Empirically, applying AML to the top-4 heads yields the best performance.

the models trained without AML and with AML, denoted as w/o AML and w/ AML, respectively. We visualize the attention maps from the 1B model, focusing on the specific layer-head pairs where AML supervision was applied. Without AML, the attention maps are notably sparse and diffuse, showing limited focus on the relevant object regions. In contrast, with AML, the attention becomes significantly sharper and more concentrated within the correct object areas. This is evident for both the `[SEG_ACT]` (adult's hand) and `[SEG_TAR]` (child) tokens, each token attending reliably to the object it is responsible for segmenting. These results demonstrate that AML enhances the MLLM's ability to allocate attention *distinctly* for each interaction-aware token, thereby strengthening role-specific segmentation.

Effect of AML on supervised and non-supervised heads. Fig. A3 compares the attention maps from (a) heads directly supervised by AML and (b) non-supervised heads within the same layer (Layer 22 of the 1B model). The supervised heads correspond to those explicitly selected for AML training, while the non-supervised heads did not receive direct supervision. Notably, we observe that the non-supervised heads also exhibit improved attention focus on the corresponding object regions, despite not being explicitly trained with AML. This indicates that the supervision signal from AML can propagate within a layer, positively influencing other heads and contributing to more consistent spatial grounding across the entire attention module.

B. Additional experimental results

In this section, we provide further experimental results supporting the impact of our dataset and architecture.

B.1. Quantitative results on RVOS benchmarks

Tab. A4 presents quantitative comparison on standard RVOS benchmarks [3, 6, 8]. Our training dataset is constructed by combining MeViS, Ref-Youtube-VOS, Ref-DAVIS, and our proposed InterRVOS-127K. ReVIOSa achieves competitive performance across all benchmarks: 46.8 $\mathcal{J}\&\mathcal{F}$ on MeViS, 71.1 on Ref-Youtube-VOS, and 76.0 on Ref-DAVIS. These results, with 4B model size, demonstrate that ReVIOSa not only excels on InterRVOS-127k but also generalizes effectively to standard RVOS benchmarks.

B.2. Comparison of training datasets

Tab. A5 compares the performance of baseline model (same setting as Tab.3 (i) in the main paper) when trained on different datasets and evaluated on the MeViS [3] benchmark,

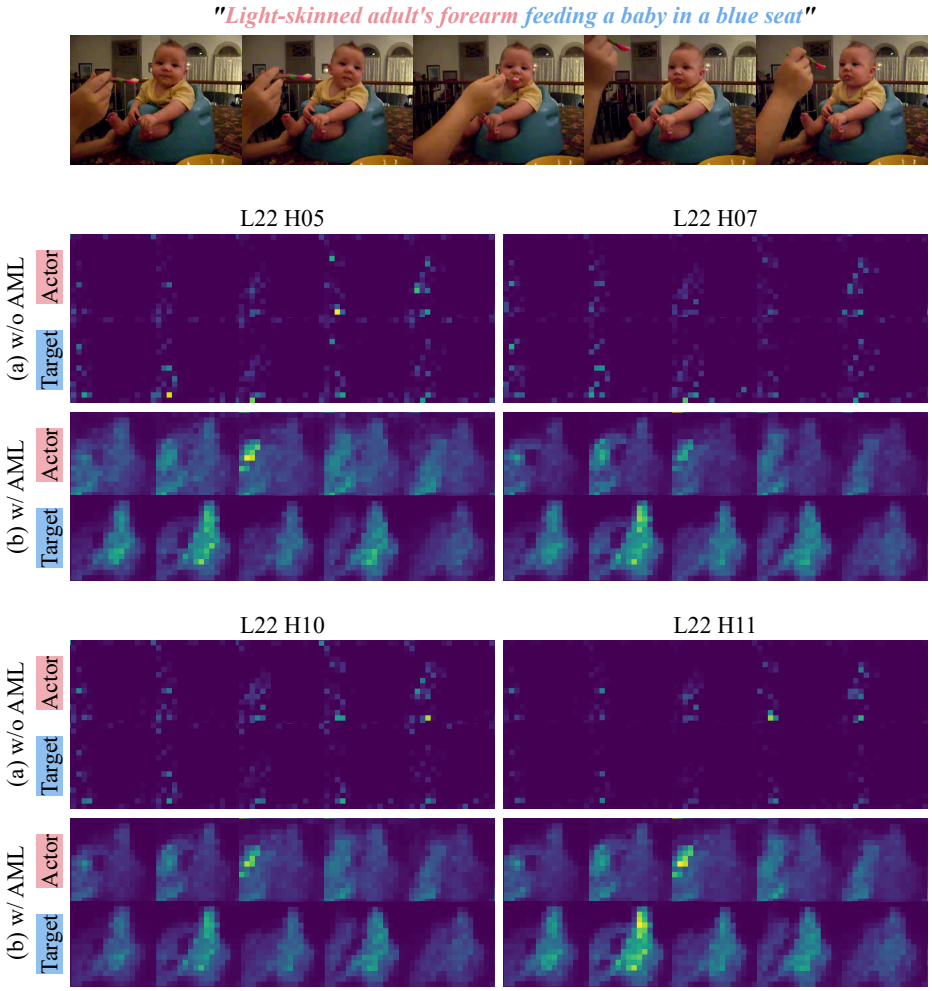


Figure A2. **Visualization of attention maps for trained layer-head pairs.** Comparison between attention maps of baseline (w/o AML) and w/ AML.

which is large enough to reliably assess model capability. Although Ref-SAV [11] is a large-scale dataset with 37K videos and 72K expressions, our subset training dataset, with only 2K videos and 28K expressions, achieves better performance. Even when controlling for sample size by matching the number of expressions, models trained on our dataset (InterRVOS-71K) outperform those trained on Ref-SAV, with the performance gap being notable in the zero-shot setting. These results validate the quality of our proposed dataset and the effectiveness of our stage-wise data annotation pipeline for capturing fine-grained information within videos.

B.3. ReVIOSa with other MLLM

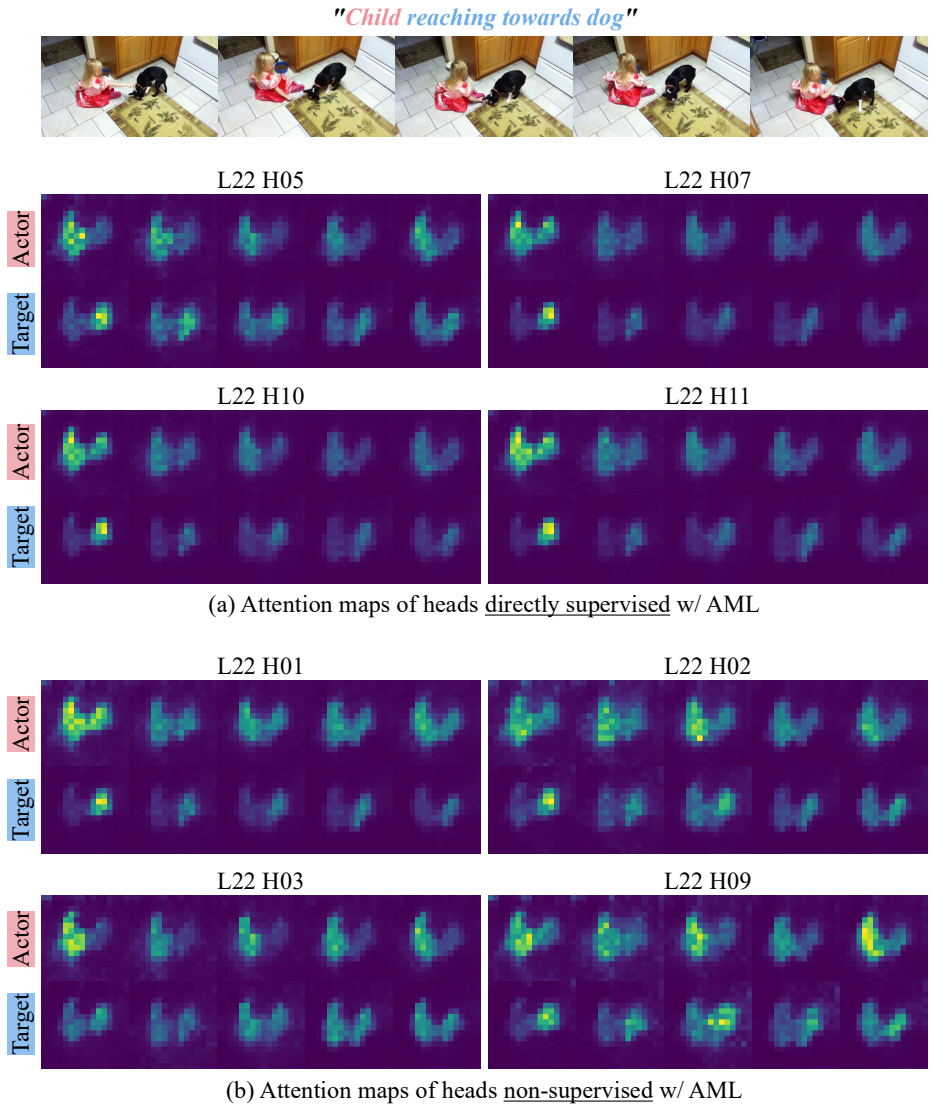
To examine generalizability, we conduct ablation using Qwen2.5-VL-3B [1] (Tab. A2, Fig. A4). ReVIOSa with Qwen2.5-VL-3B consistently improves performance across InterRVOS-Actor, InterRVOS-Target, and RVOS settings, suggesting that our architecture design is not limited to InternVL-2.5 [2].

Method	InterRVOS-Actor	InterRVOS-Target	RVOS
Baseline	64.1	-	55.0
w/ Tokens	69.5 (+1.4)	60.2 (-)	58.7 (+3.7)
w/ AML	**70.5** (+1.0)	**64.9** (+4.7)	**61.5** (+2.8)

Table A2. **Ablation with Qwen2.5-VL-3B.**

B.4. Additional qualitative results

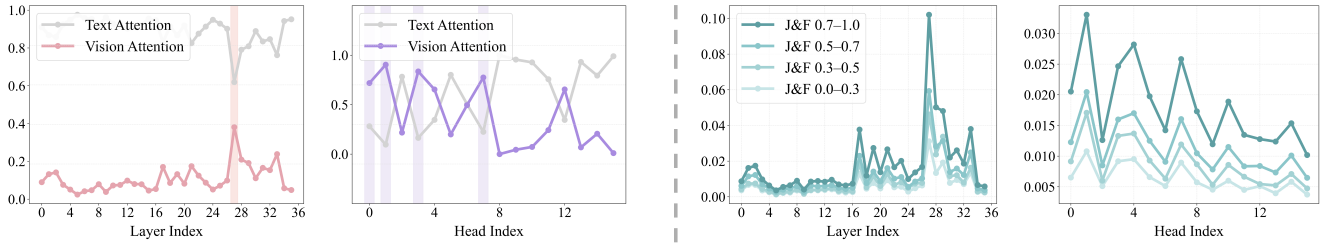
Results on InterRVOS-127K. We present qualitative results to demonstrate the effectiveness of our proposed model in handling complex, interaction-centric referring expressions. Fig. A6 compares our model (ReVIOSa) with a competitive comparison method (Sa2VA) on the proposed InterRVOS-127K dataset for the RVOS task. Across a range of challenging scenarios involving ambiguous appearance, subtle motion, and fine-grained interactions, ReVIOSa consistently achieves more accurate and temporally consistent



(a) Attention maps of heads directly supervised w/ AML

(b) Attention maps of heads non-supervised w/ AML

Figure A3. **Effect of AML on supervised vs. non-supervised heads.** Attention maps from various heads of Layer 22 of the 1B model. (a) Heads directly supervised by AML. (b) Non-supervised heads within the same layer. Notably, even without direct supervision, the non-supervised heads exhibit improved focus, suggesting that the beneficial effect of AML propagates within the layer.



(a) Layer-wise vision attentions (b) Head-wise vision attentions (Layer 27) (c) Layer/Head-wise Correlation between vision attention and J&F

Figure A4. **Qwen2.5-VL-3B Layer-Head Analysis.**

segmentation results. Notably, it exhibits strong alignment between the visual targets and the language expressions. In addition to standard referring segmentation, our model is also designed to output both actor and target objects when given interaction expressions. As illustrated in Fig. A7 and A8, the model utilizes dedicated `[SEG_ACT]`

Metrics	Descriptions
Diversity ↑	Unique predicates / Total expressions (higher = more diverse)
Singleton Ratio ↑	Predicates appearing only once / Unique predicates (higher = less repetition)
Top-10 Coverage ↓	Coverage by top 10 predicates (lower = less bias)
Top-20 Coverage ↓	Coverage by top 20 predicates (lower = less bias)

Table A3. **Summary of metrics used in dataset analysis.**

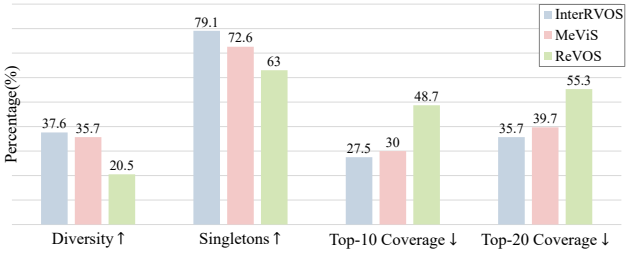


Figure A5. **Analysis on the diversity and bias of InterRVOS-127K.**

and [SEG_TAR] tokens to simultaneously localize both the actor and target described in interaction expressions.

Expression variation. We evaluate ReVIOSa's ability to adapt its segmentation according to varying referring expressions on the same video. Given different expression types, the model accurately segments the corresponding objects as specified. Fig. A11 and Fig. A12 show that ReVIOSa precisely adjusts its output based on the given expressions, correctly distinguishing the motion of each object. This validates the model's versatility across both standard RVOS and InterRVOS settings, demonstrating its fine-grained capability in understanding complex motions and disambiguating object roles.

Results on MeViS and Ref-Youtube-VOS. Fig. A9 and Fig. A10 present qualitative results on MeViS and Ref-YouTube-VOS, respectively. Across both datasets, ReVIOSa consistently captures the correct objects and produces stable mask predictions, even in scenes with multiple objects or complex motions. These examples demonstrate that our model performs robustly not only in the InterRVOS setting but also in standard RVOS scenarios.

C. Additional details of InterRVOS-127K

C.1. Distribution Analysis

We further analyzed predicate distributions using the metrics in Tab. A3. As shown in Fig. A5, compared with human-annotated datasets, InterRVOS-127K achieves the highest diversity, confirming low bias.

C.2. Data annotation pipeline

Our automatic data annotation pipeline consists of four-stage process. Among these, **Stage 1** and **Stage 3** utilize GPT-4o [5] to extract accurate object-level and interaction-level information from video contexts. In contrast, **Stage 2** and **Stage 4** focus on converting this structured information into natural language referring expressions, for which we employ the quantized version of the LLaMA 3.1 Instruct model [4].

To complement the overview in the main paper, we provide a more detailed explanation of the annotation pipeline here. Our stage-wise design enables a progressive buildup of annotation complexity, from basic object-level descriptions to more complex interaction-aware expressions.

Stage 1: Single object information. In the first stage, we focus on individual objects to obtain rich descriptions encompassing both appearance and motion attributes. We highlight a single object within the video frame and give it as an input, then GPT generates comprehensive object-centric captions that form the foundation for downstream stages. These descriptions ensure that each object is sufficiently characterized before reasoning about their interactions.

Stage 2: Single and multi-instance referring expressions. In this stage, the captions obtained from Stage 1 are reformulated into referring expressions. We handle both single object and multi-instance cases: (1) Single object expressions are generated by separating the original caption into three distinct types: appearance-only, motion-only, and combined (appearance and motion), offering finer-grained reference diversity. (2) Multi-instance expressions are created by analyzing motion similarities across objects. If multiple objects exhibit similar motion patterns, the model decides whether to merge them into a single referring expression, thereby supporting both atomic and collective object references.

Stage 3: Interaction information. In the third stage, we explore potential interactions among multiple objects within the video. Each object is annotated with an index label (e.g., [0], [1]) and fed into GPT to assess whether interactions are present. If interactions exist, we further distinguish between two types: (1) Unidirectional interactions, where a clear actor-target relationship exists (e.g., "Object [0] is leaning against object [2]"). For each pair, we generate two pseudo-captions with roles reversed (e.g., "Object [2] is being leaned on by object [0]") and extract structured actor-target mappings. (2) Bidirectional interactions, where objects participate equally (e.g., "Object [0] and object [1] are standing together with arms around each other"). In such cases, only the object pair involved is extracted without role assignment. This stage is crucial for capturing the relational structure between entities and building a pool of interaction data that reflects both directionality and symmetry.

Stage 4: Interaction-aware referring expressions. In the final stage, we convert structured interaction information from Stage 3 into rich referring expressions. Starting from GPT-generated index-based captions (e.g., "Ob-

Methods	MeViS	Ref-Youtube-VOS	Ref-DAVIS
LISA-7B [7]	39.4	54.3	64.8
LISA-13B [7]	37.9	54.4	66.0
TrackGPT-7B [12]	40.1	56.4	63.2
TrackGPT-13B [12]	41.2	59.5	66.5
VISA-7B [10]	43.5	61.5	69.4
VISA-13B [10]	44.5	63.0	70.4
Sa2VA-4B [11]	<u>46.2</u>	**71.3**	<u>73.7</u>
ReVIOSa-4B	**46.8**	<u>71.1</u>	**76.0**

Table A4. **Quantitative results on RVOS benchmarks.**

Dataset	Setting	Ref-SAV (Videos 37k / Exps. 72K)	InterRVOS-28K (Videos 2k / Exps. 28K)	InterRVOS-71K (Videos 5K / Exps. 71K)
MeViS valid	Joint Training	46.8	**48.5**	<u>47.1</u>
	Zero-shot	32.8	<u>40.2</u>	**41.8**
MeViS valid_u	Joint Training	53.0	<u>54.6</u>	**54.8**
	Zero-shot	40.1	<u>50.1</u>	**50.5**

Table A5. **Effectiveness of InterRVOS-127K.** Despite using fewer samples, models trained on InterRVOS-28K and InterRVOS-71K outperform the Ref-SAV dataset [11] (72K) on MeViS [3] benchmark in both the joint training setting (with MeViS [3] train set) and the zero-shot setting (with only InterRVOS-28K and InterRVOS-71K train sets). This highlights the superior data efficiency and interaction-centric supervision quality of the InterRVOS-127K dataset.

ject `[0]` is leaning against object `[2]`"), we inject class and appearance description for each object obtained from stage 2 to produce semantically enriched expressions. This yields two levels of interaction captions: (1) Class-level, using coarse object category labels (2) Appearance-level, incorporating visual attributes from earlier stages.

Throughout the entire data annotation pipeline, the InterRVOS-127K dataset evolves into a diverse and large-scale resource that simultaneously provides rich descriptions of object interactions, ranging from simple to highly detailed expressions.

C.3. Additional examples of InterRVOS-127K

Fig. A14 and Fig. A15 present additional examples from the InterRVOS-127K dataset. Our dataset covers a broad range of referring expressions, including challenging cases like multi-object references and motion-only descriptions, as well as varying levels of granularity from class-level to fine-grained appearance-based expressions. It also includes interaction-focused expressions that clearly distinguish actor and target roles. The examples illustrate multiple objects within a single video and their relationships, highlighting the dataset's ability to capture object-level interactions in complex scenes.

C.4. Video clip extraction procedure

The InterRVOS-127K dataset is constructed using source videos from the VidOR dataset [9], which contains a large number of long-form videos, many exceeding 1,000 frames in length. To generate more diverse and effective video clips for referring video object segmentation, we apply a systematic clip extraction strategy. Specifically, each original source video is divided into non-overlapping temporal bins of 1,000 frames. From these, we select only the first and last bins to increase the likelihood of capturing distinct scenes or transitions within a single video. Within each selected bin, we extract only the first 500 frames to form a video clip. This approach allows us to generate a wide range of video segments while ensuring sufficient temporal context and diverse scenes required for RVOS. As a result, we obtain high-quality video clips that are both temporally coherent and suitable for dense language grounding and interaction modeling.

C.5. Dataset statistics

The overall statistics of the InterRVOS-127K dataset are presented in Fig. A16.

The word frequency distribution (a) reveals that commonly used terms such as *object*, *person*, *child*, *side*, *position*, and *right* frequently appear in the referring expressions. This indicates that the dataset captures not only static appearance information but also emphasizes spatial relations and interactive contexts involving everyday entities. In terms of temporal characteristics, (b) shows that most videos fall within the 10 to 20 second range, providing suffi-

cient temporal context for modeling object-level dynamics. Additionally, (c) illustrates the distribution of video frames: the training set mostly consists of 500 frames, while the validation set is composed of shorter clips with frame counts aligned in increments of 5.

The dataset also exhibits significant linguistic density and visual complexity. As shown in (d), most videos are annotated with 5 to 20 referring expressions, peaking at the 10 to 15 range, which enables dense language grounding for each clip. Moreover, (e) indicates that a large portion of videos contain 0 to 5 annotated objects, with a smaller but meaningful subset containing more than 5. This diversity in object count allows the dataset to cover a broad range of scene complexities, from simple to highly interactive scenarios. Collectively, these statistics confirm that the InterRVOS-127K dataset is well-suited for advancing research in referring video object segmentation and interaction-centric video understanding.

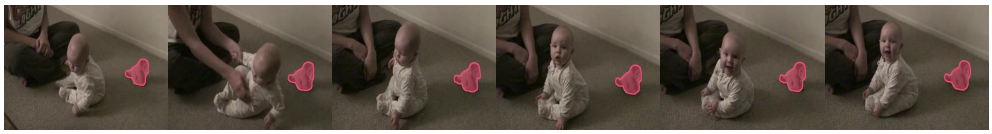
Furthermore, (f), (g), and (h) provide an overall interaction-focused statistics within InterRVOS-127K. In (f), we observe that approximately 65% of videos contain at least one interaction-based referring expression, indicating that interaction scenarios are prevalent throughout the dataset. (g) further illustrates the distribution of the number of interaction expressions per video, and (h) shows the number of objects involved in each interaction; while most interactions involve two objects, a notable 20.3% involve three, suggesting a considerable portion of the dataset covers more complex, multi-object interactions.

D. Failure cases and future works

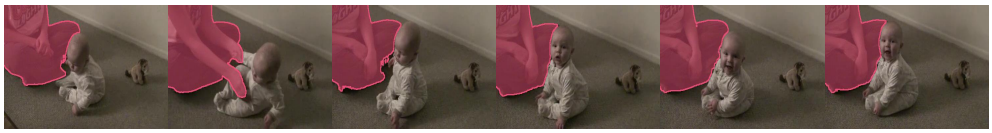
Despite its strong interaction modeling capability, ReVIOSa exhibits a few remaining limitations, as illustrated in Fig. A13. One issue arises when the objects to segment appear at relatively small scales within the frames, making precise localization and segmentation challenging (top). Another difficulty occurs when interactions involve a large number of entities, such as cases with three or more objects to segment, the model faces difficulties in generating consistent mask tracks across frames (bottom).

In future work, we aim to extend ReVIOSa to handle role switching over longer video sequences, enabling the model to track how actor-target relationships change overtime. Another promising direction is part-level segmentation, allowing the model to segment only the specific part of an object involved in the interaction, for example, segmenting just the right arm of a person when the referring expression is *"the right arm holding the table"*. Advancing in these directions would support richer interaction reasoning and further push the boundaries of RVOS beyond its current capabilities.

User: Please segment "*the stationary object on the floor*".



ReVIOSa: Sure, [SEG_ACT].



Sa2VA: Sure, [SEG].

User: Please segment "*the object kneeling on the ground, slightly adjusting its position behind a green barrier*".

ReVIOSa: Sure, [SEG_ACT].

Sa2VA: Sure, [SEG].

User: Please segment "*the seated object, mostly still with occasional slight shifts in the chair*".

ReVIOSa: Sure, [SEG_ACT].

Sa2VA: Sure, [SEG].

User: Please segment "*the large white dog with fluffy, dense fur that remains mostly stationary while slightly shifting its position, occasionally turning its head*".

ReVIOSa: Sure, [SEG_ACT].

Sa2VA: Sure, [SEG].

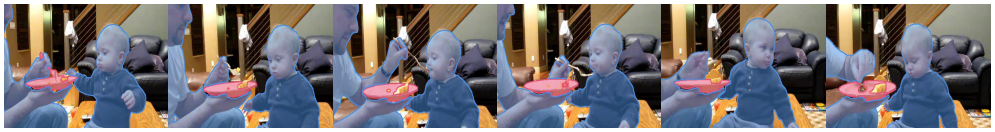
Figure A6. **Qualitative results on InterRVOS-127K: RVOS Setting.** Qualitative comparisons between our model (ReVIOSa) and a strong comparison model (Sa2VA) on the proposed InterRVOS-127K dataset for the RVOS task. ReVIOSa consistently produces more accurate and temporally consistent segmentation masks, especially in challenging scenarios involving fine-grained interactions, appearance ambiguity, or motion.

User: Please segment "*black dog with a white marking being led by person in a dark jacket and gloves*".

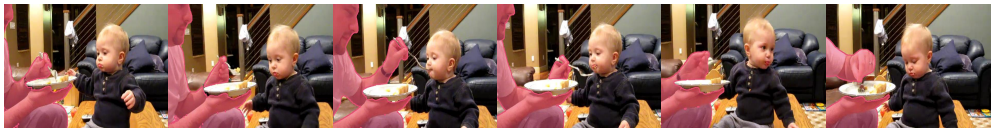
ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

Sa2VA: Sure, [SEG].

User: Please segment "*dish feeding person and child*".



ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].



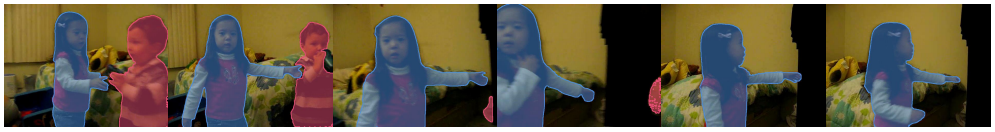
Sa2VA: Sure, [SEG].

User: Please segment "*person walking a dog*".

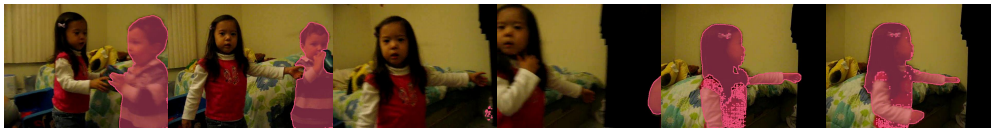
ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].

Sa2VA: Sure, [SEG].

User: Please segment "*striped child having his blue cup taken by a pink-shirt child*".



ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].



Sa2VA: Sure, [SEG].

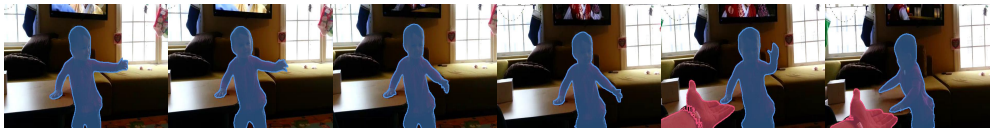
Figure A7. **Qualitative results on InterRVOS-127K: InterRVOS setting.** Qualitative results of ReVIOSa on InterRVOS-127K. By using interaction-aware special tokens (`[SEG_ACT]` and `[SEG_TAR]`) and AML, our model separately segments actor (pink) and target (blue) objects, effectively capturing their distinct roles and behaviors within interactions. This demonstrates ReVIOSa's ability to understand asymmetric inter-object relationships and model the dynamics that define video interactions.

User: Please segment "*child reaching out to man*".

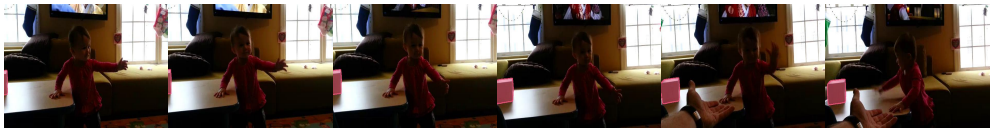
ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].

Sa2VA: Sure, [SEG].

User: Please segment "*hand reaching towards child*".

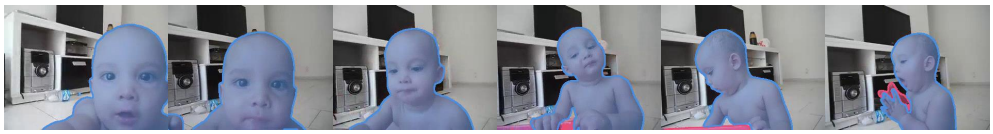


ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].



Sa2VA: Sure, [SEG].

User: Please segment "*glossy black toy with blue element held by young baby*".



ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].

Sa2VA: Sure, [SEG].

User: Please segment "*pink bottle being held by baby with giraffe diaper*".

ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

Sa2VA: Sure, [SEG].

Figure A8. **Qualitative results on InterRVOS-127K: InterRVOS setting.** Qualitative results of ReVIOSa on InterRVOS-127K. By using interaction-aware special tokens (`[SEG_ACT]` and `[SEG_TAR]`) and AML, our model separately segments actor (pink) and target (blue) objects, effectively capturing their distinct roles and behaviors within interactions. This demonstrates ReVIOSa's ability to understand asymmetric inter-object relationships and model the dynamics that define video interactions.

User: Please segment "*dog circling on green plate*".

ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*The bear that had been pinned down by the other bear*".

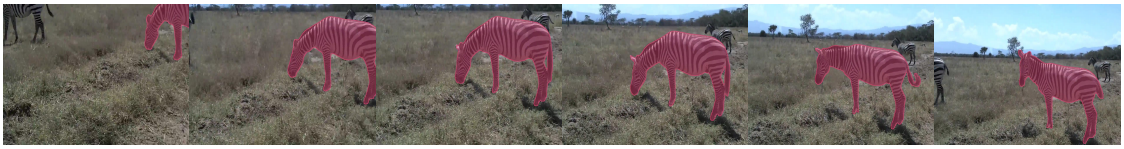
ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*cat lying on the ground* ".

ReVIOSa: Sure, it is [SEG_ACT].

Figure A9. **Qualitative results on MeViS dataset.**

User: Please segment "*a black and white zebra is on the right eating the grass*".



ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*a rabbit is weighed on a weighing machine*".

ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*a blue and white colored surfboard in the right hand of dark blue swim suit*".

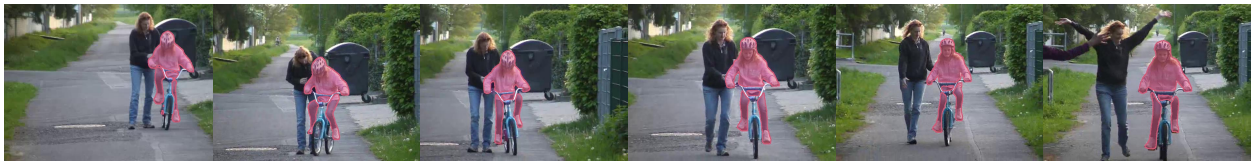
ReVIOSa: Sure, it is [SEG_ACT].

Figure A10. **Qualitative results on Ref-Youtube-VOS dataset.**

User: Please segment "*The adult wearing a black jacket and blue jeans with shoulder-length hair, walking alongside a child on a bicycle and then raising their arms as the child cycles away independently*".

ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*the child in a pink jacket and white helmet riding a blue bicycle and moving forward while maintaining balance*".



ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*Adult supporting child and bicycle*".

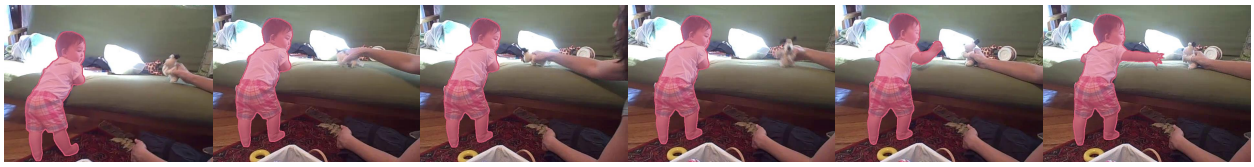
ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

User: Please segment "*Bicycle being ridden by child with support from adult*".

ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

Figure A11. **Qualitative results with diverse expression variations.** As the given expression changes, ReVIOSa precisely localizes the corresponding objects accordingly, demonstrating robust and fine-grained understanding across both standard RVOS and InterRVOS setting.

User: Please segment "*Object leaning forward and slightly shifting weight*".

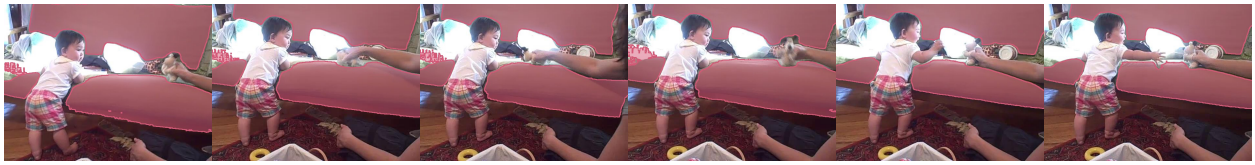


ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*Object moving gently back and forth, shifting slightly in position as it is held and manipulated*".

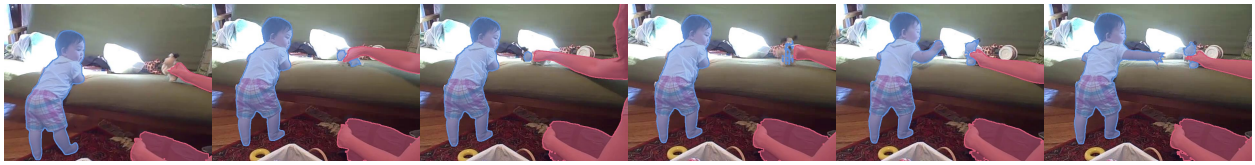
ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*The green, rectangular sofa with a smooth texture that remains stationary in the room*".



ReVIOSa: Sure, it is [SEG_ACT].

User: Please segment "*adult handing toy to child*".

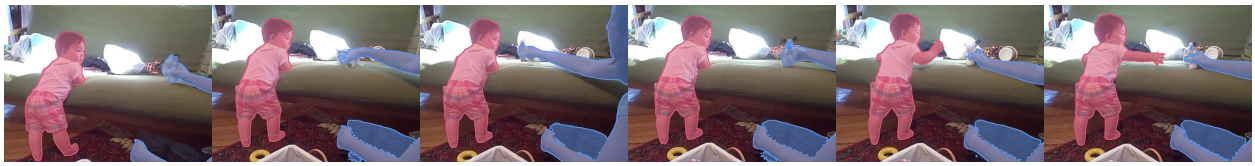


ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

User: Please segment "*Toy being handed by adult to child*".

ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

User: Please segment "*Child receiving toy from adult*".



ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

Figure A12. **Qualitative results with diverse expression variations.** As the given expression changes, ReVIOSa precisely localizes the corresponding objects accordingly, demonstrating robust and fine-grained understanding across both standard RVOS and InterRVOS setting.

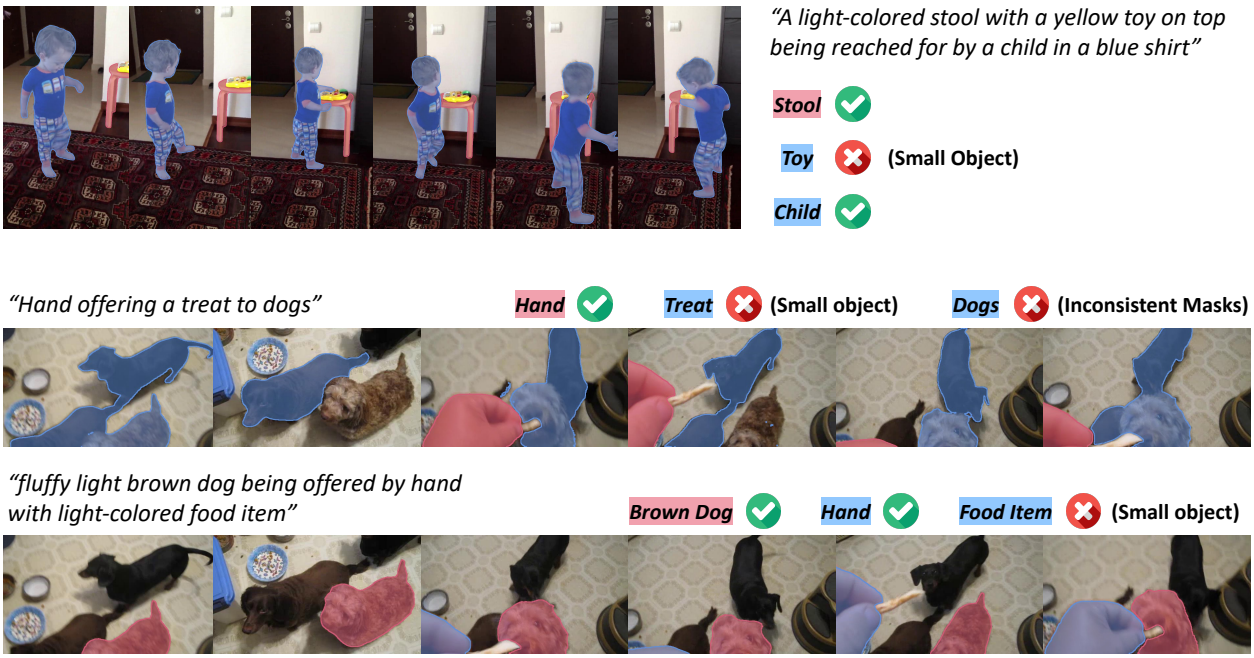


Figure A13. **Failure cases.**

[Referring Expressions]

Object [0]

"The person wearing a plaid shirt and gloves reaching toward and unwrapping the foil-wrapped object with their hands"

Object [1]

"The object moving around the space, handling a metal bowl wrapped in foil before walking towards a table and setting it down"

"The person wearing a dark blue patterned long-sleeve shirt and jeans"

Object [2]

"Object standing in place with a hand in pocket"

"Adult wearing a red long-sleeved shirt, blue jeans, and white shoes"

Objects [0], [1]

"People working together to unwrap foil"

"The one in plaid shirt and the one in dark blue patterned shirt working together to unwrap foil"

[Actor-Target Expressions]

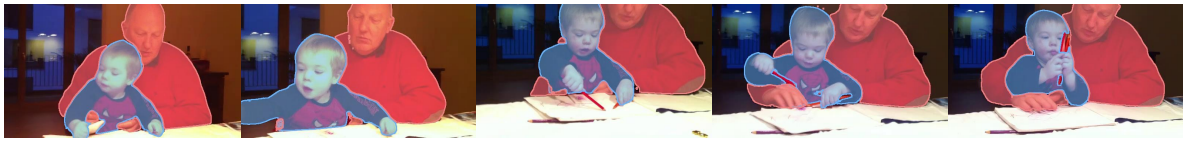
Actor [0] / Target [1]

"person handing to person"

"person in plaid shirt handing to person in dark blue pattered long-sleeve shirt"

Actor [1] / Target [0]

"Person receiving item from person"



[Referring Expressions]

Object [0]

"Young child with light brown hair and a red shirt featuring a superhero logo"

"Object moving arms back and forth while drawing"

Object [1]

"The man wearing a bright red sweater, with short hair and a focused expression, interacting with a child"

Objects [0], [1]

"The child and the man interacting at the table"

[Actor-Target Expressions]

Actor [0] / Target [1]

"child being assisted by man in drawing"

"young child with red superhero shirt being assisted by man in bright red sweater in drawing"

Actor [1] / Target [0]

"Man helping a child"

"Man in bright red sweater helping child with superhero shirt"

Figure A14. **Examples of InterRVOS-127K.**

[Referring Expressions]

Object [0]

"Person wearing a white T-shirt with a logo on the back and red pants, standing with hands on hips, moving towards the open car trunk, bending slightly forward, and returning to a standing position facing the car"

"Object standing with hands on hips, moving towards the open car trunk, bending slightly forward, and returning to a standing position facing the car"

Object [1]

"Adult in a light-colored shirt, dark knee-length shorts, and sneakers with red and white detailing"

Object [2]

"Object shifting position slightly and gesturing with a hand as it moves towards the back of a car"

"Adult in a white short-sleeved shirt, dark shorts, and dark shoes, shifting position slightly and gesturing with their hand as they move towards the back of a car"

Object [3]

"The sporty white car with various decals and a prominent spoiler that remains stationary with its rear compartment opened and inspected"

Objects [0], [2]

"People moving around a car"

[Actor-Target Expressions]

Actor [0] / Target [3]

"person working on car"

"person with logo on back working on sporty white car with decals"

Actor [3] / Target [0]

"Car being worked on by person"

Actor [1] / Target [2], [3]

"Person listening to person and looking at car"

"Man in light-colored shirt listening to man in white shirt and looking at sporty white car"

Actor [2] / Target [1], [3]

"Person explaining to person and car"

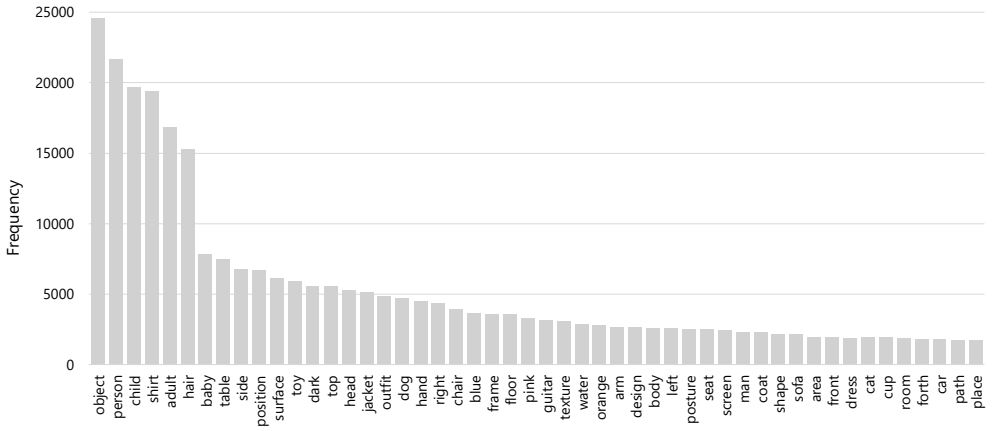
"Adult in white shirt explaining to adult in light-colored shirt and sporty white car"

Actor [3] / Target [1], [2]

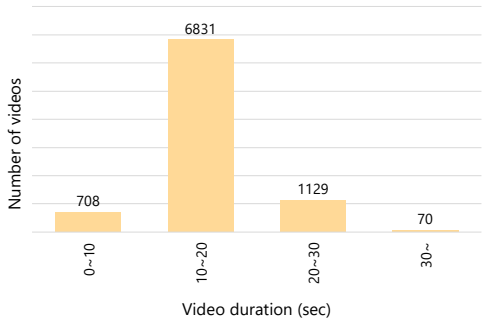
"Car being discussed by people"

"Sporty white car with decals being discussed by light-shirt person and white-shirt person"

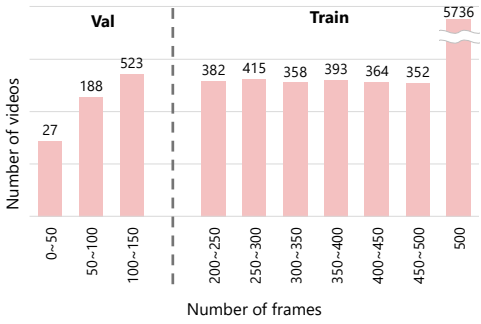
Figure A15. **Examples of InterRVOS-127K.**



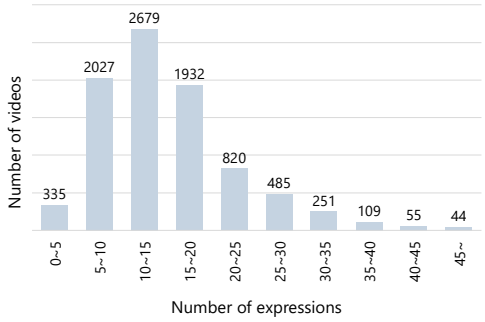
(a) Frequency distribution of the Top-50 most frequent words



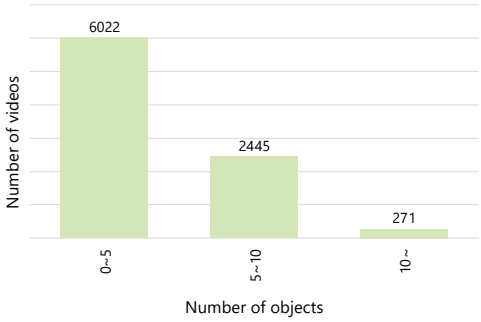
(b) Distribution of video duration



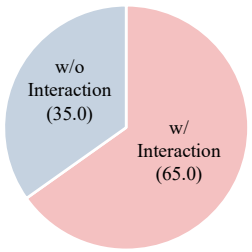
(c) Distribution of video frames



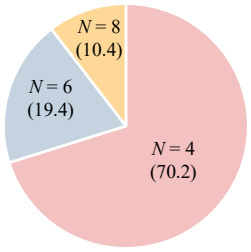
(d) Number of expressions per video



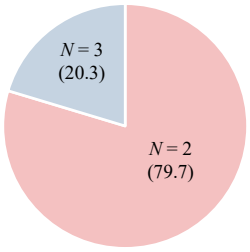
(e) Number of objects per video



(f) Videos with interaction expressions



(g) Interaction expressions per video



(h) Objects engaged in one interaction

Figure A16. **Overall statistics of InterRVOS-127K.**

Stage 1 : Single object information (GPT-4o)

<task>
You are given a video where specific objects are highlighted. Your task is to describe only the highlighted object, focusing on both its visual appearance and how it moves or changes position throughout the video.
</task>
<objectives>
1. Provide a **localized caption** that describes:
  - The visual **appearance** (color, shape, texture, category, etc.) of the highlighted object.
  - The object's **motion** or **spatial movement** (e.g., moving left, jumping, rotating).
2. Do not mention any other objects that are not highlighted.
3. Use only the information that can be **visually confirmed** from the video. **Do not infer or assume anything** that is not clearly visible (e.g., names of people, unobservable intent or unseen background).
4. **Do not refer to the red highlight, colored contour, or any visual marking used to identify the object.** Focus only on the object's inherent visual and behavioral properties.
5. Use clear, concise language that reflects what is visually and spatially observable from the highlighted object only.
6. The object's motion description must refer to **the same highlighted object** whose appearance you just described. Do not describe movement of unrelated objects, background elements, or the overall scene.
7. If the highlighted object is stationary or only slightly moving, describe that accurately. Do not fabricate or exaggerate movement based on nearby motion.
</objectives>
<inputDetails>
- The input is a short video clip containing multiple objects.
- One or more objects are highlighted using a **colored contour around their boundary**.
- The video is designed to preserve the **object's appearance** and provide visual cues for its **motion** across frames.
- Focus only on the object with the **colored boundary**, but do **not** describe the boundary or outline itself in your output.
</inputDetails>
<objectClass>
- The object class is `"{kwargs["obj_class"]}"`.
- Use this information only to support your understanding of what kind of object to describe.
- However, you must describe **the object that is visually highlighted** in the video (e.g., marked with a red boundary or mask).
- If there are multiple objects of the same class in the scene, **focus solely on the highlighted one**, even if others appear more salient or central.
</objectClass>
<outputFormat>
Provide **two distinct sentences** in a single paragraph form:
1. Describe what the object looks like (e.g., "A small brown dog with curly fur and a blue collar.")
2. Describe how the object moves or behaves in the video (e.g., "It runs from left to right across the grassy field, occasionally looking back.")
Avoid describing things that cannot be visually confirmed from the video.
</outputFormat>

Figure A17. **Stage 1: Input prompts to GPT-4o.** We provide GPT-4o with preprocessed video frames in which objects are highlighted using labels and colored masks. This stage aims to extract localized information for each object, including both appearance and motion attributes.

Stage 2 : Single and multi-instance referring expressions (LLaMA-70B)

Stage 2-1 : Single object referring expressions

```
"role": "system",
"content": (
```

You are an assistant that generates **referring captions** for a single object in a video.
You will be given two descriptions of the object:
- An **appearance** description (what it looks like)
- A **motion** description (how it moves or changes position)
Your task is to convert these descriptions into **natural referring expressions**, while preserving as much information as possible.
Generate three outputs:
1. A caption that combines both appearance and motion (key: 'all')
2. A caption that uses only the motion (key: 'motion')
3. A caption that uses only the appearance (key: 'appearance')
IMPORTANT RULES:
- Rewrite each caption as a **referring expression**, not a full sentence.
- Use singular form only. Never use plural expressions like 'they' or 'their'. Assume the object is a single entity.
- Do not use the word ‘figure’. Use an alternative. Especially for the ‘motion’ description, use terms like ‘object’ or others that do not imply appearance.
- **Do not omit details** from the input descriptions. Keep the meaning and key attributes intact.
- Rephrase **only as needed** to make the output sound like a natural referring phrase.
- Do NOT add new information or hallucinate.
- Avoid phrases like 'The object is' or 'This is'.
Output must be in the following strict JSON format: {

```
  "all": "<caption combining appearance and motion>",
  "motion": "<caption using only motion>",
  "appearance": "<caption using only appearance>"
}
)

"role": "user",
"content": (
f"appearance_caption: {gpt_appearance_caption},
f"motion_caption: {gpt_motion_caption}
```

Please generate the referring captions in the specified JSON format, following the rules above.

```
)
```

Figure A18. **Stage 2 (Single-object case): Input prompts to LLaMA.** Using the object-level descriptions generated in Stage 1, we prompt LLaMA to produce diverse referring expressions. For single-object cases, we decompose the description into three types: appearance-only, motion-only, and combined expressions.

Stage 2 : Single and multi-instance referring expressions (LLaMA-70B)

Stage 2-2 : Multi-instance referring expressions

```
"role": "system",
"content": (
```

You are an assistant that analyzes multiple objects in a video based on their motion captions.
Your task is to determine whether any objects can be grouped together into a single referring caption, based on whether they:
1. Belong to the **similar object class** (e.g., person, hand, cup, phone)
2. Share **semantically similar motion behaviors**
3. Are **describing the same primary object** (not just interacting with the same object)
IMPORTANT RULES:
- For each object, only consider the **main object being described** in its motion caption.
Do NOT merge objects that describe **different entities**, even if similar objects are mentioned in the background.
- For example, 'A hand holding a phone' and 'A phone moving near the face' describe different main subjects (hand vs. phone) and should NOT be merged.
- If the motion captions indicate that the objects are **stationary** or show **no meaningful movement**, then do NOT merge them.
Only merge objects that share clear and active motion behaviors (e.g., crawling, lowering, walking, waving, spinning, moving around, sitting at a couch, watching TV).
Output Format (JSON only):
- 'merged': 'YES' or 'NO'
- 'merged_objects': List of object IDs that were merged (or null if no merge)
- 'merged_caption': Referring caption describing the shared motion (or null if no merge)
Stylistic Rules for merged_caption:
- Use explicit object class (e.g., 'the people', 'the cups') — do not use pronouns like 'they'.
- Write a **referring-style phrase**, not an explanatory sentence. Example: 'People walking side by side', not 'The people are walking...'
- Your output must be valid JSON. No extra text or commentary.

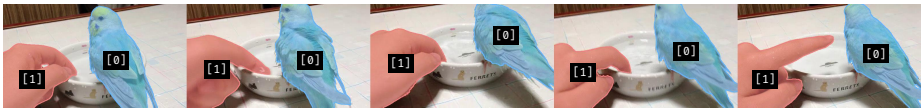
```
)

"role": "user",
"content": (
f"obj_captions: {video_objs_caption_dict}
```

Please determine if any objects can be merged based on object class and motion similarity and return the result in the specified JSON format.

```
)
```

Figure A19. **Stage 2 (Multi-instance case): Input prompts to LLaMA.** For videos containing multiple objects with similar motion, we prompt LLaMA to determine whether they should be merged into a single referring expression. The decision is made based on motion similarity.

Stage 3 : Interaction information (GPT-4o)

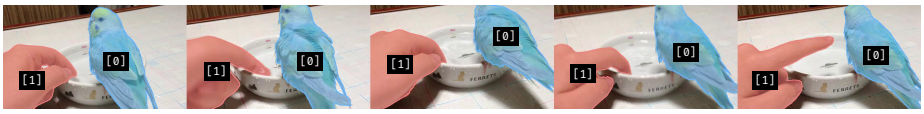


<task>
You are given a video in which multiple labeled objects appear. Your task is to identify any visible interaction between the labeled objects, determine the type and direction of interaction, and describe it appropriately.
</task>
<objectives>
1. Determine whether any interaction is visually observable between the labeled objects.
2. If yes, classify the interaction as:
  - "bidirectional" (e.g., mutual interaction like "[2] and [3] are dancing together")
  - "unidirectional" (e.g., directional interaction like "[0] is handing something to [1]")
3. For each interaction:
  - If bidirectional → provide **one sentence** describing the mutual interaction.
  - If unidirectional → provide **two sentences**:
    - One where the **initiator** is the subject
    - One where the **receiver** is the subject (in passive form)
  - **Include all objects that are directly or indirectly involved in the interaction in the `object_pair` list.**
  - **If the interaction is `unidirectional`, provide one sentence for each object in `object_pair`, using that object as the grammatical subject.**
    - For example, if `object_pair` is ["[0]", "[1]", "[7]"], there should be three sentences:
      - One with [0] as the subject
      - One with [1] as the subject
      - One with [7] as the subject
4. Interactions involving more than two objects (e.g., [0], [1], [2]) should be described as a group if they jointly participate in the same action.
5. Always refer to objects using their exact labels like "[1]", "[2]", etc.
6. Only describe interactions that are visually verifiable—do not infer hidden intentions, emotions, or relationships.
</objectives>
<inputDetails>
- The input video contains labeled objects with the following identifiers:
 `{kwargs["valid_obj_ids"]}`
- These are the only valid object labels. You must not use or invent any other object identifiers.
- Each object is highlighted with a colored outline.
</inputDetails>
<additionalInput>
The following object categories are provided as prior knowledge:
`obj_categories = {kwargs["obj_categories"]}`
These categories may guide your understanding of plausible interactions, but your final decisions must rely strictly on visual evidence.
</additionalInput>

(continue)

Figure A20. **Stage 3: Input prompts to GPT-4o.** We provide GPT-4o with preprocessed frames highlighting all objects with labels and colored masks. This stage focuses on detecting interactions between objects and generating detailed descriptions of their relationships.

Stage 3 : Interaction information (GPT-4o)



<reasoningSteps>
Step-by-step reasoning:
1. Consider only the labeled objects: {`kwargs["valid_obj_ids"]`}
2. Do not assume the existence of any other object labels (e.g., [0], [3] are invalid).
3. Examine all valid pairs and groups of the provided objects.
4. For each candidate interaction:
  a. Observe their motion, spatial alignment, and relative timing.
  b. If interaction occurs:
    i. Classify it as bidirectional or unidirectional.
    ii. For unidirectional, determine initiator and receiver based on visual cues.
  c. After writing the descriptions:
  - Ensure that every object in `object_pair` appears as the **grammatical subject** of at least one sentence.
5. Construct appropriate descriptions accordingly.
6. If no interactions are observed, return interaction = "NO".
</reasoningSteps>
<outputFormat>

```
{{
  "interaction": "YES" or "NO",
  "interactions": [
    {{
      "object_pair": ["[1]", "[2]"],
      "type": "bidirectional",
      "descriptions": [
        "Object [1] and object [2] are shaking hands."
      ]
    }},
    {{
      "object_pair": ["[8]", "[2]"],
      "type": "unidirectional",
      "descriptions": [
        "Object [8] is pointing at object [2].",
        "Object [2] is being pointed at by object [8]."
      ]
    }}
  ] or None
}}
```

</outputFormat>
<selfCheck>
Before finalizing your output:
- Double-check that every object mentioned in the descriptions is present in the `object_pair`.
- Double-check that each object in the object_pair appears as the grammatical **subject** in at least one sentence.
</selfCheck>

Figure A21. **Stage 3 : Input prompts to GPT-4o.** We provide GPT-4o with preprocessed frames highlighting all objects with labels and colored masks. This stage focuses on detecting interactions between objects and generating detailed descriptions of their relationships.

Stage 4 : Interaction referring expressions (LLaMA-70B)

Stage 4-1 : Bidirectional

```
"role": "system",
"content": (
```

You are an assistant that generates **referring captions** describing interactions between objects in a video.
Input:
- '`obj_captions`': a dictionary of object IDs mapped to their appearance descriptions
- '`interaction_description`': a natural language sentence involving object IDs (e.g., 'Object [0] and object [1] are sparring.')
Your task is to generate two types of referring captions by replacing the object references in the `interaction_description` with natural expressions that identify them:
1. `class_level`: Use high-level object class names only (e.g., 'person', 'child')
2. `appearance_level`: Use short, distinguishing appearance descriptions (not full captions, just enough to tell them apart)
Output Format:
- Return a dictionary in JSON format with the following two keys:
  - `class_level`
  - `appearance_level`
Stylistic Rules:
- Referring captions must be concise and natural phrases (not explanatory sentences)
- Do NOT write full explanatory sentences like 'The A is doing B with the C'
Instead, write expressions like 'A doing B with C' or 'The one in red jacket sparring with the one in white shirt'
- You may omit verbs like 'is' or 'are' to keep the sentence minimal and referential in style
- Do NOT use pronouns like 'they' or 'their'.
- Do NOT write full sentences like 'The people are...'. Instead, write: 'People sparring with each other'.
- If both objects belong to the same class, you may use a plural collective form like 'People', 'Children', etc.
- The appearance-level caption should reflect just enough visual detail from `obj_captions` to distinguish the two objects naturally.

```
)

"role": "user",
"content": (
f"obj_captions: {obj_captions}
f"interaction_description: {interaction_description}
"Please return your response as a JSON dictionary containing the referring captions."
)
```

Figure A22. **Stage 4 (Bidirectional case): Input prompts to LLaMA.** We prompt LLaMA using interaction-level descriptions generated in Stage 3. Appearance and class information from Stage 2 are injected into each entity, indicated by labeled placeholders (e.g., `[0]`).

Stage 4 : Interaction referring expressions (LLaMA-70B)

Stage 4-2 : Unidirectional

`"role": "system",`
`"content": (`
You are an assistant that generates **referring captions** describing interactions between objects in a video.
Input:
- `obj_captions`: a dictionary of object IDs mapped to their appearance descriptions
- `interaction_description`: a natural language sentence involving object IDs (e.g., 'Object [0] is hugging object [1]')
- `subject_id`: the ID of the object performing the action
- `object_id`': the ID of the object receiving the action
Your task is to generate two types of referring captions:
1. `class_level`: Use object class names only (e.g., 'person', 'cup', 'bear')
2. `appearance_level`: Use short, distinguishing appearance descriptions (not the full description — just enough to distinguish the object)
Output Format:
- Return a JSON dictionary with keys:
 - `class_level`
 - `appearance_level`
Important Rules:
- Carefully reflect the subject (agent) and object (recipient) roles as provided in `subject_id` and `object_id`.
- Do NOT follow the order in the sentence — follow the subject-object mapping explicitly.
- The referring captions must be short, descriptive, and in the form of **natural referring phrases** — not full explanatory sentences.\n"
- Avoid structures like 'The A is doing B to the C'. Instead, use expressions like:
  - 'Parrot watching at person'
  - 'Person feeding a rabbit'
- Do NOT use pronouns like 'they' or 'their'.
- The appearance-level caption should reflect just enough visual detail from obj_captions to distinguish the two objects naturally.
`)`

`"role": "user",`
`"content": (`
`f"obj_captions: {obj_captions}`
`f"interaction_description: {interaction_description}`
`f"subject_id: {subject_id}`
`f"object_id: {object_id}`
Please return your response as a JSON dictionary containing the referring captions.
Do **not** include any other description, explanation, or formatting — just the JSON dictionary.
`)`

Figure A23. **Stage 4 (Unidirectional case): Input prompts to LLaMA.** In cases where the interaction is classified as *unidirectional*, LLaMA additionally predicts actor object and target object identifiers. This enables us to assign distinct segmentation mask tracks to each role.

References

[1] 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, 2025. 3

[2] 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. 3

[3] 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. 2, 6

[4] 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. 5

[5] 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. 5

[6] 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. 2

[7] 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*, pages 9579–9589, 2024. 6

[8] 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. 2

[9] 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. 6

[10] 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. 6

[11] 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. 3, 6

[12] Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *arXiv preprint arXiv:2312.17448*, 2023. 6