Interaction-Aware Referring Video Object Segmentation

Woojeong Jin, Seongchan Kim, and Seungryong Kim
KAIST AI
{wooj0216, star._.overflow, seungryong.kim}@kaist.ac.kr

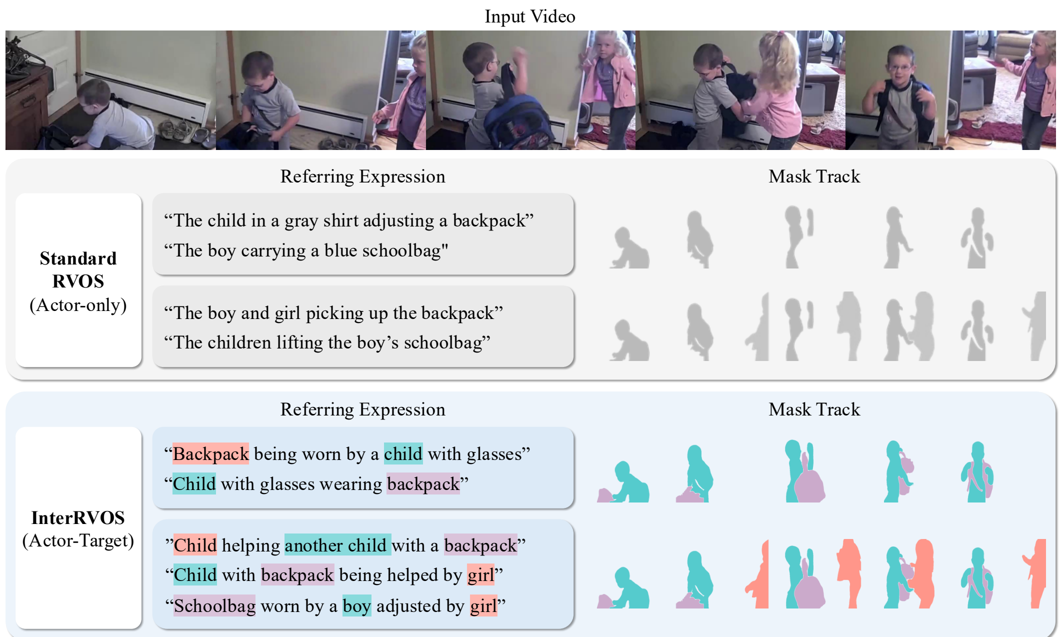


Figure 1: **Teaser.** Existing datasets (Ding et al., 2023; Seo et al., 2020; Khoreva et al., 2019; Yan et al., 2024) for standard referring video object segmentation focused solely on *actor-only* settings and primarily included referring expressions describing single or multiple objects with similar actions, and thus they lack sufficient examples involving object interactions that require reasoning over both the *actor* and the *target*. To address this gap, we introduce a new task, **interaction-aware referring video object segmentation (InterRVOS)**, which centers on interaction-aware referring expressions and considers distinct segmentation masks for both the actor and its interacting target objects.

Abstract

Referring video object segmentation aims to segment the object in a video corresponding to a given natural language expression. While prior works have explored various referring scenarios, including motion-centric or multi-instance expressions, most approaches still focus on localizing a single target object in isolation. However, in comprehensive video understanding, an object’s role is often defined by its interactions with other entities, which are largely overlooked in existing datasets and models. In this work, we introduce **interaction-aware referring video object segmentation (InterRVOS)**, a new task that requires segmenting both actor and target entities involved in an interaction. Each interaction is described through a pair of complementary expressions from different semantic perspectives, enabling fine-grained modeling of inter-object relationships. To tackle this task, we propose **InterRVOS-8K**, the large-scale and automatically constructed dataset containing diverse interaction-aware expressions with corresponding masks, including challenging cases such as motion-only multi-instance expressions. We also present a baseline architecture, **ReVIOSa**, designed to handle actor-target segmentation from a single expression, achieving strong performance in both standard and interaction-focused settings. Furthermore, we introduce an actor-target-aware evaluation setting that enables a more targeted assessment of interaction understanding. Experimental results demonstrate that our approach outperforms prior methods in modeling complex object interactions for referring video object segmentation task, establishing a strong foundation for future research in interaction-centric video understanding. Our project page is available at: `https://cvlab-kaist.github.io/InterRVOS`.

1 INTRODUCTION

Referring video object segmentation (RVOS) aims to segment the object in a video that corresponds to a given natural language expression. Existing RVOS benchmarks (Seo et al., 2020; Gavrilyuk et al., 2018; Yan et al., 2024; Khoreva et al., 2019; Ding et al., 2023) and methods (Zhou et al., 2024; Wang et al., 2023; Yuan et al., 2025; Oh et al., 2020; Ding et al., 2023; Liu et al., 2024; Wang et al., 2024) have primarily focused on aligning a single expression with a single target object, predicting its segmentation mask across video frames. While recent efforts (Ding et al., 2023; Yan et al., 2024) have begun addressing more complex referring scenarios—such as motion-centric or multi-instance expressions—they still largely assume that each expression refers to objects exhibiting similar behavior, without explicitly identifying the roles of involved entities or modeling their interactions.

In contrast, real-world video understanding often requires detailed understanding about inter-object dynamics, where an object's identity and relevance are shaped not only by its appearance or movement, but also by *how it interacts with others*. Despite the critical role of such interactions between objects, they remain largely underexplored in current RVOS datasets and architectures.

Consider the interaction of "A child helping another child with a backpack", as illustrated in Figure 1. This interaction can be expressed through three complementary expressions, each highlighting a different role: "Child helping another child with a backpack" (actor-focused; typically covered in conventional RVOS), "Child with backpack being helped by girl" (target-focused), and "Schoolbag worn by a boy adjusted by girl" (target-focused). These expressions offer distinct yet semantically aligned perspectives on the same interaction, each highlighting a different participating entity. A model must not only understand these distinct roles but also accurately segment all entities mentioned across the expressions.

Building on this, we formulate **interaction-aware referring video object segmentation (InterRVOS)** as a task that segments entities involved in a single interaction, each described from distinct semantic roles. Specifically, we construct paired expressions that describe the same interaction from different perspectives—such as the actor object and the target objects participating in the single interaction. The model is expected to generate accurate segmentation masks for all entities mentioned across these expressions, enabling a comprehensive understanding of the interaction. We define an interaction-aware expression as one that (1) describes an observable interaction in the video, and (2) explicitly identifies the actor and the target involved.

To support this task, we introduce **InterRVOS-8K**, a large-scale and automatically annotated dataset specifically designed for interaction-aware referring video object segmentation. Our dataset facilitates the learning of temporal dynamics and inter-object distinctions by including a rich set of interaction expressions. In addition to covering diverse interaction types, InterRVOS-8K also incorporates challenging cases such as multi-instance expressions and motion-only descriptions. This diversity enables models trained on InterRVOS-8K to generalize better to complex, real-world video understanding scenarios. Compared to prior datasets (Seo et al., 2020; Ding et al., 2023; Yuan et al., 2025; Gavrilyuk et al., 2018; Yan et al., 2024; Khoreva et al., 2019), which are often small in scale or centered on single-object descriptions, InterRVOS-8K serves as a more comprehensive dataset for modeling interaction reasoning in video, as compared in Table 1.

We also present a baseline architecture, **ReVIOSa**, tailored for this task. Our model not only performs referring segmentation, but also extends the standard setting by enabling segmentation of both actor and target from a single interaction expression. The model learns to disentangle roles between entities and to generate accurate segmentation masks for each, enhancing both instance-level discrimination and overall interaction understanding. The experiments on ReVIOSa demonstrate that our proposed

Datasets	Annotation	Size	Single	Multiple	Actor-Target Interact.
A2D Sentence (Gavrilyuk et al., 2018)	Manual	6.6K	✓	✗	✗
J-HMDB Sentence (Gavrilyuk et al., 2018)	Manual	0.9K	✓	✗	✗
Ref-DAVIS (Khoreva et al., 2019)	Manual	1.5K	✓	✗	✗
Ref-Youtube-VOS (Seo et al., 2020)	Manual	15K	✓	✗	✗
MeViS (Ding et al., 2023)	Manual	28K	✓	✓	✗
ReVOS (Yan et al., 2024)	Manual	25K	✓	✓	✗
Ref-SAV (Yuan et al., 2025)	Automatic	72K	✓	✗	✗
InterRVOS-8K	Automatic	127K	✓	✓	✓

Table 1: **Comparison of existing RVOS datasets and the InterRVOS-8K dataset.** Unlike existing datasets, the InterRVOS-8K dataset supports all three types of referring expressions—single object, multiple objects, and actor-target interaction—and is the largest to date (127K samples). It is the first to explicitly annotate interactions between actor and targets.

architecture achieves better interaction understanding compared to previous methods (Zhou et al., 2024; Yuan et al., 2025; Wu et al., 2022; Ding et al., 2023; Wang et al., 2023). We further introduce an actor-target-aware evaluation setting, which follows the standard RVOS evaluation protocol but is applied only to interaction-aware samples that inherently form expression pairs reflecting complementary actor-target roles within the same interaction. This evaluation enables a more targeted assessment of a model's ability to understand diverse interaction semantics and distinguish between entities involved in an interaction.

In summary, our contributions are as follows:

- We introduce InterRVOS, a new task that extends conventional RVOS by requiring segmentation of both actor and target entities involved in a given pair of interaction-aware expressions.
- We present InterRVOS-8K, the first large-scale and automatically annotated dataset specifically tailored for this task, covering diverse and challenging scenarios including multi-instance expressions, motion-only expressions, and temporally complex scenes.
- We develop a baseline architecture ReVIOSa that accurately segments both actor and target roles from a single interaction-aware expression.
- We propose an actor-target-aware evaluation setting that enables more rigorous and targeted assessment of interaction understanding.

2 Related work

Referring video object segmentation. RVOS aims to segment a target object in a video based on a natural language expression. Early studies primarily focused on multi-modal fusion between language and visual features (Gavrilyuk et al., 2018; Oh et al., 2020; Ding et al., 2021; Botach et al., 2022; Wu et al., 2022; Miao et al., 2023), operating mostly in single-frame or single-object settings with limited temporal modeling. To address the lack of temporal reasoning, MeViS (Ding et al., 2023) introduced a relatively large-scale and more challenging dataset, including motion-only expressions and multi-instance samples. This has encouraged models to incorporate motion-aware and spatio-temporal reasoning to better track and understand objects across time. More recently, ReVOS Yan et al. (2024) extends the RVOS task toward reasoning, aiming to model high-level understanding by converting expressions into inference-based questions. These developments have improved the scope of RVOS by targeting motion and reasoning, but they still fall short in modeling inter-object interaction, which plays a fundamental role in many real-world scenarios.

Meanwhile, approaches such as DsHmp (Zhou et al., 2024) and SOC (Wang et al., 2023) leverage text-encoder-based frameworks with lightweight architectures, showing strong performance without relying on large-scale multimodal decoders. More recent methods employ multi-modal large language models (MLLMs) (Liu et al., 2023) with special tokens (e.g., `[SEG]`) to prompt segmentation (Yuan et al., 2025; Liu et al., 2024; Wang et al., 2024; Bai et al., 2024). While these approaches demonstrate generalization and zero-shot capabilities, they do not explicitly model dynamic relations between multiple objects—an essential component for resolving complex or role-based referring expressions.

In terms of benchmarks, most existing datasets—such as Refer-Youtube-VOS (Oh et al., 2020), A2D-Sentences (Gavrilyuk et al., 2018), and MeViS (Ding et al., 2023)—are either limited in scale or focused on single-object annotations. Even Ref-SAV (Yuan et al., 2025) remains single object focused and lacks challenging cases such as motion-centric or multi-instance expressions, as well as scenarios involving explicit inter-object relationships. As a result, current datasets still underrepresent complex relational expressions such as "the person holding the bag" or "the ball passed by the player," which require models to understand structured interactions between entities.

To overcome these limitations, we introduce a new interaction-centric dataset and task that explicitly model actor-target relations. It includes paired expressions describing the same interaction from different viewpoints (e.g., actor and target), enabling fine-grained alignment between language and video. Our approach extends RVOS beyond isolated object localization and supports large-scale relational video understanding.

Video object interaction. Understanding object interactions is a fundamental challenge in video understanding, as many events and activities are defined by the dynamic relations between multiple entities. Prior work has introduced datasets with structured annotations of visual relationships, including VidOR (Shang et al., 2019), ImageNet-VidVRD (Shang et al., 2017), and Action Genome (Ji et al., 2020), which capture triplet-based interactions (actor, predicate, target) across time. These datasets enable models to learn how entities influence each other, such as "person riding bicycle" or"dog chasing ball", which are essential for high-level activity reasoning. Recent efforts have expanded these ideas into temporally grounded visual situations. For example, STAR (Wu et al., 2024) offer fine-grained temporal and causal relation annotations across multiple interacting agents, facilitating a deeper understanding of scene dynamics. The MOMA dataset (Fan et al., 2021) goes further by modeling complex manipulation actions and agent-tool-object interactions in procedural contexts. These datasets and models demonstrate the importance of reasoning about inter-object roles, especially when resolving ambiguous or role-based queries.

Despite these advancements, most RVOS models still do not fully incorporate interaction reasoning and instead assume isolated object descriptions, overlooking cases where relational cues are crucial (e.g., "the man holding the red cup"). In this work, we address this gap by introducing an interaction-centric RVOS dataset and a baseline model, aiming to integrate relational understanding into the segmentation process.

3 INTERRVOS-8K DATASET

Despite recent advances in RVOS, interactions between objects remain underexplored in existing datasets. Existing datasets like Ref-SAV (Yuan et al., 2025) lack the diversity required to represent the wide range of events and interactions that naturally occur in real-world videos. We introduce **InterRVOS-8K**, a large-scale InterRVOS dataset focused on interaction-aware expressions. InterRVOS-8K is designed to facilitate deeper video understanding by emphasizing *actor-target interactions*, a key yet previously overlooked component in the RVOS domain. InterRVOS-8K dataset is build upon VidOR (Thomee et al., 2016), and is constructed through a stage-wise automated annotation pipeline that leverages GPT-4o (Hurst et al., 2024) and LLaMA-70B (Grattafiori et al., 2024) to generate and verify high-quality captions.

3.1 DATA ANNOTATION PIPELINE

To generate diverse interaction-aware referring expressions, we design a stage-wise annotation pipeline consisting of four main stages. This progressive structure enables us to gradually build up from basic object-level descriptions to more complex interaction-level expressions. The overall stages of data annotation pipeline is illustrated in Figure 2.

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 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.

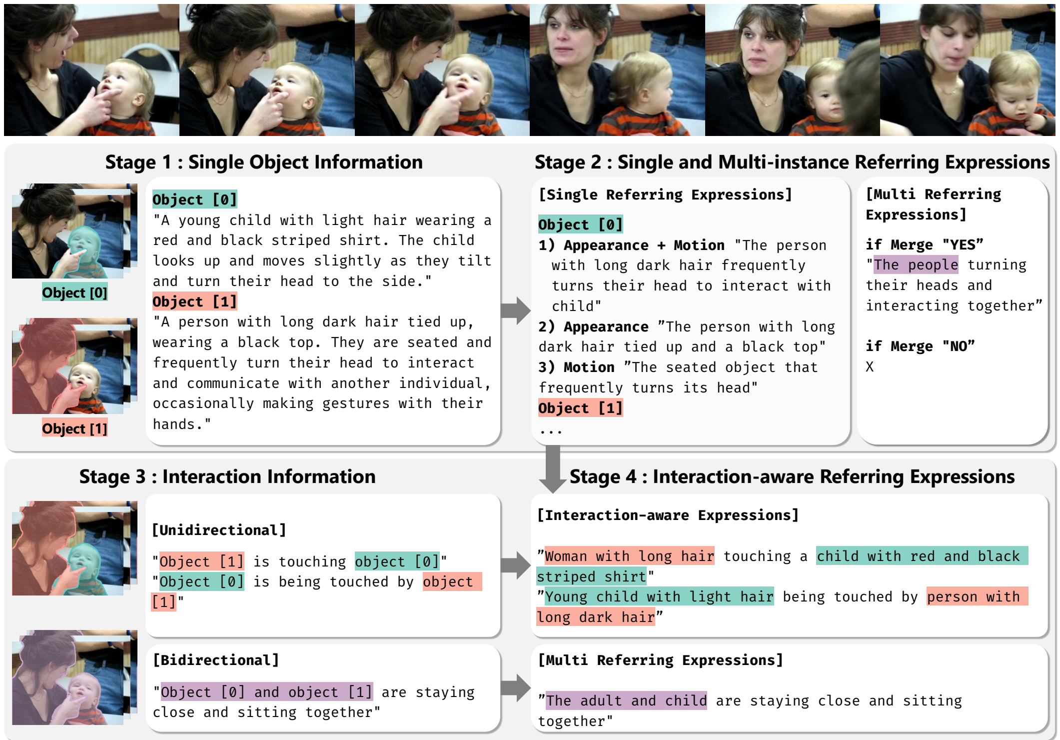


Figure 2: **Data annotation pipeline.** Our proposed automatic data annotation pipeline constructs referring expressions for single, multi-object, and interaction scenarios in four stages, which extracts object appearance and motion, detects inter-object interactions, and generates detailed expressions grounded in both visual properties and interaction context.

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 decide 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., "Object `[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.

Datasets	Video	Object	Expression	Object/Video	Actor-Target Interaction
A2D Sentence (Gavrilyuk et al., 2018)	3,782	4,825	6,656	1.28	-
J-HMDB Sentence (Gavrilyuk et al., 2018)	928	928	928	1	-
Ref-DAVIS (Khoreva et al., 2019)	90	205	1,544	2.27	-
Ref-Youtube-VOS (Seo et al., 2020)	3,978	7,451	15,009	1.86	-
MeViS (Ding et al., 2023)	2,006	8,171	28,570	<u>4.28</u>	-
ReVOS (Yan et al., 2024)	1,042	5,535	35,074	**5.31**	-
Ref-SAV (Yuan et al., 2025)	**37,311**	**72,509**	<u>72,509</u>	1.94	-
InterRVOS-8K (Ours)	<u>8,738</u>	<u>35,247</u>	**127,314**	4.03	**17,682**

Table 2: **Comparison of various RVOS datasets.** Our newly proposed **InterRVOS-8K** offers the largest number of referring expressions and a high object-per-video ratio, enabling richer and more diverse visual grounding across complex scenes compared to existing benchmarks. Unlike existing datasets, InterRVOS-8K also provides interaction-aware referring expressions that explicitly distinguish between actor and target roles, enabling fine-grained understanding of visual interactions.

Throughout the entire data annotation pipeline, the InterRVOS-8K dataset evolves into a diverse and large-scale resource that simultaneously provides rich descriptions of object interactions, ranging from simple to highly detailed expressions.

3.2 Data analysis and statistics

We perform a statistical comparison between existing RVOS datasets and InterRVOS-8K. As shown in Table 2, our dataset not only surpasses most of the previous datasets in overall scale, but also demonstrates a significant advantage in terms of the number of expressions, and annotated objects per video. Given that a large number of objects and diverse referring expressions are essential for modeling complex scenarios in the RVOS domain (Ding et al., 2023), these statistics validate that InterRVOS-8K is well-constructed to serve as a large-scale dataset.

Collectively, these statistics underscore that InterRVOS-8K is not only positioned as a large-scale dataset in the RVOS domain, but also uniquely targets the underexplored yet crucial aspect of *actor-target object interaction*. In particular, it includes interaction-aware referring expressions that explicitly distinguish between actor and target roles within interactions, enabling more precise and detailed relational understanding of complex video scenes. This makes our dataset particularly valuable for models that aim to understand and segment interactions between multiple objects in video. We provide additional details about the InterRVOS-8K dataset in Appendix C. This includes word frequency, video duration, the number of expressions and objects, as well as interaction-centric statistics such as the number of interactions and the number of objects involved in each interaction.

Figure 3 presents examples from the InterRVOS-8K dataset. As shown, our dataset includes a wide range of referring expressions, covering both challenging cases such as multi-object references and motion-only descriptions, as well as a diverse spectrum of expression granularity—from simple class-level descriptions to fine-grained appearance-based references. In addition to conventional referring expressions, InterRVOS-8K explicitly incorporates interaction-focused expressions that distinguish between actor and target roles. The examples also demonstrate the presence of multiple objects within a single video and highlight the relationships between them, confirming that our dataset effectively captures object-level interactions in complex visual scenes. Additional examples are provided in Appendix B.

3.3 Evaluation set

We construct the InterRVOS-8K evaluation set by selecting a subset of videos from the training split of the VidOR dataset (Shang et al., 2019). To ensure a clear separation between training and evaluation, all videos in the evaluation set are strictly disjoint from those used during training. The evaluation set comprises 738 videos and 5,126 referring expressions. Given that our dataset emphasizes object interactions, we apply a filtering process to increase the proportion of interaction-centric expressions while reducing the occurrence of simple single-object references in the evaluation set. Further details on the construction of the training and evaluation sets are provided in the Appendix C.

[Referring Expressions]

Object [0]

"The object that eagerly moves its head towards the ice cream cone, occasionally pushing closer as it tries to reach the treat"

Object [1]

"Object standing close to a person, eagerly licking an ice cream cone"

Object [2]

"The object moving slightly up and down as it presents to a dog"

Objects [0], [1]

"The dogs moving to reach the ice cream cone"

[Actor-Target Expressions]

Actor [2] / Target [0], [1]

"Person holding for black and brown dog and medium-sized dog"

"Pale hand holding ice cream for black and brown dog and tan and white dog"

Actor [0] / Target [2]

"Dog licking a hand"

"Black and brown dog licking a pale hand with a light-colored ice cream cone"

Actor [1] / Target [2]

"Dog licking person"

"Tan dog licking hand with ice cream"

Figure 3: **Examples from the InterRVOS-8K dataset.** InterRVOS-8K dataset consists of diverse referring expressions, including multi-object, motion-only, class-level, and appearance-based descriptions. Notably, the dataset captures actor-target interactions and relationships among multiple objects within a video.

4 ReVIOSa Architecture

As our proposed dataset InterRVOS-8K emphasizes a detailed understanding of object interactions and diverse motion dynamics, we present **ReVIOSa** (**Re**ferring **V**ideo **I**nteraction-aware **O**bject **S**egment**a**tion), a baseline architecture tailored for this purpose. Unlike prior RVOS approaches that typically segment only the actor referred to in the language expression, ReVIOSa is designed to jointly reason about and segment both the actor object and the target object, especially in cases involving unidirectional interactions. This enables a richer interpretation of referring expressions by explicitly modeling the relational context between interacting entities.

4.1 Overview

The overall architecture of ReVIOSa is illustrated in Figure 4. It builds upon SAM2 (Ravi et al., 2024) and a LLaVA-based (Liu et al., 2023) multimodal large language model (MLLM), combining strong visual understanding with language-guided segmentation. Given a text prompt and an input video, the model extracts visual features using both the SAM2 encoder and the MLLM's visual encoder, while language features are obtained from the LoRA-enhanced (Hu et al., 2022) MLLM. During training, only the LoRA modules and the SAM2 decoder are updated, while all other components remain frozen. We introduce two special segmentation tokens to enable interaction-aware mask generation. These tokens act as prompts for the SAM2 decoder, which produces distinct segmentation masks for the actor and the target, enabling explicit modeling of object interactions in referring expressions.

4.2 Interaction-aware special tokens

To enable actor-target aware segmentation, we extend the segmentation prompting scheme with interaction-aware special tokens. Specifically, we define `[SEG_ACT]` to represent the actor object and `[SEG_TAR]` to represent the target object involved in an interaction. The model dynamically determines which tokens to emit based on the nature of the referring expression. When the expression

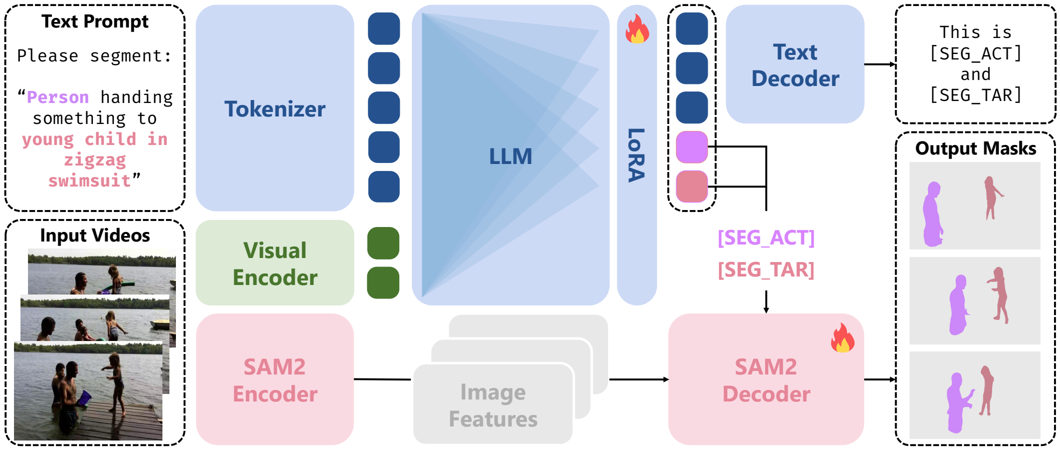


Figure 4: **Our baseline architecture.** Our model retains the [SEG_ACT] token to segment the referred subject and introduces a new [SEG_TAR] token to additionally segment the interacting object. This enables explicit role-aware segmentation for interaction-aware referring expressions.

describes a single entity or a bidirectional interaction, the model outputs only the [SEG_ACT] token. In contrast, for expressions that describe a unidirectional interaction (e.g., "the person pushing the cart"), the model generates both [SEG_ACT] and [SEG_TAR] tokens to represent the actor ("the person") and the target ("the cart"), respectively.

These special tokens are embedded into the LLM output sequence and serve as spatial-temporal prompts for the SAM2 decoder. The hidden state associated with each token is projected and used to guide the segmentation mask generation. The model is trained to associate each token with its corresponding mask, enabling it to distinguish between the roles of entities involved in the interaction. The overall training objective loss formulation are pixel-wise cross-entropy loss and dice loss applied to the predicted masks. Depending on whether the given referring expression describes an interaction, the loss is computed on either one or both of the masks generated by [SEG_ACT] and [SEG_TAR].

4.3 Actor-target joint segmentation

ReVIOSa is built to support a actor-target joint segmentation framework that extends the conventional RVOS formulation. During training, even when only one of the entities is required for evaluation, the model is supervised to generate both [SEG_ACT] and [SEG_TAR] tokens and their corresponding masks whenever the expression implies an interaction. This allows the model to learn to reason about both the acting actor and the affected target in a unified manner.

At inference time, for standard RVOS benchmarks that evaluate only the segmentation of the referred actor, we use the mask associated with the [SEG_ACT] token for scoring. Nevertheless, the model is inherently capable of producing masks for both entities when the expression involves interaction, enabling richer, more contextualized segmentation. We argue that by training the model to solve this more challenging task—i.e., understanding and segmenting both roles in an interaction—it gains a deeper comprehension of the video content and referring expression, which ultimately leads to improved performance even under traditional RVOS settings.

5 Experiments

In this section, we present experimental results to evaluate the effectiveness of our proposed baseline approach, ReVIOSa, as well as the impact of our newly constructed interaction-centric InterRVOS-8K dataset. We demonstrate the significance of both the model and the dataset through comprehensive comparisons and ablations. All experiments follow the standard evaluation protocol used in RVOS (Oh et al., 2020; Ding et al., 2023), where the main metric is the average of region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ denoted as $\mathcal{J}\&\mathcal{F}$. The implementation details are described in Appendix A.

Dataset	Setting	Ref-SAV (Videos 37k / Exps. 72K)	InterRVOS-2K (Videos 2k / Exps. 28K)	InterRVOS-5K (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 3: **Effectiveness of InterRVOS-8K.** Despite using fewer samples, models trained on InterRVOS-2K and InterRVOS-5K outperform the Ref-SAV dataset (Yuan et al., 2025) (37K) on MeViS (Ding et al., 2023) benchmark in both the joint training setting (with MeViS (Ding et al., 2023) train set) and the zero-shot setting (with only InterRVOS-8K train set). This highlights the superior data efficiency and interaction-centric supervision quality of the InterRVOS-8K dataset.

	Training Dataset	Architecture	Single (Appear. + Motion)	Single (Appear.)	Single (Motion)	Multi	Interaction	Overall
(a)	MeViS	Baseline	53.0	54.0	35.1	62.5	<u>61.6</u>	50.4
	Ref-SAV		53.4	53.6	22.1	32.5	37.9	41.6
	InterRVOS-5K		<u>60.6</u>	**60.5**	<u>43.0</u>	<u>67.2</u>	60.6	<u>56.4</u>
(b)	InterRVOS-5K	ReVIOSa	**61.0**	<u>59.7</u>	**44.6**	**70.0**	**65.7**	**57.6**

Table 4: **Performance analysis by caption types across different training datasets.** Models trained on subset InterRVOS-5K with architecture ReVIOSa significantly outperform those trained on MeViS (Ding et al., 2023) or Ref-SAV (Yuan et al., 2025) across all expression types, especially for motion-only and interaction expressions.

5.1 SIGNIFICANCE AND EFFECTIVENESS OF DATASET

Comparison of training datasets on MeViS benchmark. Table 3 compares the performance of the Sa2VA (Yuan et al., 2025) baseline when trained on different datasets and evaluated on the MeViS (Ding et al., 2023) benchmark. Although Ref-SAV (Yuan et al., 2025) 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 the sample size with same expression number, models trained on our dataset outperform those trained on Ref-SAV. The gap is especially notable in the zero-shot setting, where the model is evaluated on MeViS without having seen any MeViS samples during training. This indicates that Ref-SAV, while large, is limited by its single-object-centric design. In contrast, our dataset, which is automatically constructed to be diverse and interaction-aware, provides more effective supervision for video understanding tasks. To further demonstrate the effectiveness of our dataset and architecture, we provide additional experimental results in Appendix D.

Analysis on the type of captions. Table 4 presents a performance analysis by caption type across different training datasets—MeViS (Ding et al., 2023), Ref-SAV (Yuan et al., 2025), and InterRVOS-5K—evaluated on the InterRVOS-8K test set. (a) presents results when the same baseline model, Sa2VA (Yuan et al., 2025), is trained on different datasets and evaluated on our dataset's evaluation set, which can be separated into various type as illustrated in the table. We observe that training on Ref-SAV (Yuan et al., 2025) leads to lower performance than even training on MeViS (Ding et al., 2023), despite Ref-SAV's larger scale. In contrast, training on our subset InterRVOS-5K significantly improves performance across all caption types, including relatively easier appearance-based expressions. Notably, our dataset yields the most substantial gains on challenging categories such as multi-instance expressions and motion-only expressions, demonstrating its strength in modeling complex video dynamics. However, performance on interaction expressions is slightly lower compared to other categories, suggesting that even with high-quality data, the model architecture must be equipped to effectively learn and represent inter-object relationships. By comparing Table 4 (a) and (b), we further confirm that dataset quality alone is insufficient for handling interaction-centric expressions. As shown in (b), incorporating architectural components specifically designed for reasoning over interactions leads to further improvements, highlighting the complementary roles of both data and model design in advancing video understanding.

Methods	Referring			Actor-Target			Overall		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Referformer (Wu et al., 2022)	49.4	50.8	50.1	57.8	58.8	58.3	51.8	53.0	52.4
LMPM (Ding et al., 2023)	42.8	46.2	44.5	49.5	53.0	51.2	44.7	48.1	46.4
Sa2VA-1B (Yuan et al., 2025)	50.0	53.5	51.7	57.2	60.6	58.9	52.0	55.5	53.8
Sa2VA-4B (Yuan et al., 2025)	<u>53.8</u>	56.8	55.3	65.4	67.8	66.6	57.1	59.9	58.5
ReVIOSa-1B	<u>53.8</u>	<u>57.7</u>	<u>55.8</u>	**68.3**	**71.3**	**69.8**	<u>57.9</u>	**61.6**	<u>59.7</u>
ReVIOSa-4B	**54.9**	**57.9**	**56.4**	<u>68.0</u>	<u>70.6</u>	<u>69.3</u>	**58.6**	<u>61.5</u>	**60.0**

Table 5: **Quantitative results on InterRVOS-8K dataset.** ReVIOSa achieves state-of-the-art performance on both standard referring expressions (Referring) and interaction-centric expressions (Actor-Target), the latter of which are not covered by existing RVOS datasets.

Methods	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
(a) Only `[SEG]`	52.0	55.5	53.8
(b) Independent `[SEG_ACT]` `[SEG_TAR]`	52.4	56.4	54.4
(c) Joint (Ours) `[SEG_ACT]` `[SEG_TAR]`	**57.9**	**61.6**	**59.7**

Table 6: **Ablation study on ReVIOSa architecture.** Joint modeling of actor and target object segmentation leads to the best performance, demonstrating the benefit of learning interactions between entities.

5.2 Results of ReVIOSa

Quantitative results. Table 5 reports the quantitative performance under three evaluation categories: Referring, Actor-Target, and Overall. Among them, the Actor-Target segmentation setting represents our newly proposed evaluation protocol, which measures performance only on interaction expressions that contain annotated actor-target pairs. Since each interaction expression in the InterRVOS-8K dataset is constructed as a pair with clearly defined actor and targets, this setting enables evaluation of the model's ability to handle dual-role interactions. The Referring setting evaluates on standard expressions without actor-target pairing, while Overall includes all samples. Across all three settings, our proposed model, ReVIOSa, consistently achieves the highest performance. Notably, the ReVIOSa-1B model, despite having fewer parameters, outperforms the larger Sa2VA-4B (Yuan et al., 2025) model. This demonstrates not only the effectiveness of our architectural design in modeling object interactions, but also its strength in broader video understanding tasks.

Ablation studies. To validate the effectiveness of our actor-target modeling strategy, we conduct ablation studies on model of size 1B focused on the use of special tokens: `[SEG_ACT]` and `[SEG_TAR]`. Table 6 summarizes the results. In setting (a), we use a single `[SEG]` token without distinguishing between actor and target roles. In setting (b), we introduce both `[SEG_ACT]` and `[SEG_TAR]`, and train the model with separate supervision: actor mask supervision is applied only when the output token is `[SEG_ACT]`, and likewise for `[SEG_TAR]`. Setting (c) corresponds to our proposed method, where the model generates both `[SEG_ACT]` and `[SEG_TAR]` as outputs simultaneously, enabling the model to learn interaction semantics between actor and target roles jointly. The results demonstrate that explicitly modeling both roles improves segmentation performance and promotes better understanding of inter-object relationships.

Qualitative results. Figure 5 compares qualitative results under two inference settings: (a) standard referring and (b) interaction-aware referring, which is aware of both actor and target roles. In (a), our model accurately segments non-salient target objects, showing strong alignment between segmentation mask and language. In (b), it successfully segments both actor object and target objects, demonstrating actor-target awareness and the ability to model interaction semantics. Additional qualitative results can be found in Appendix E.

Figure 5: **Qualitative results.** Qualitative comparison between standard referring and interaction-aware referring. The model shows strong language-segmentation alignment in standard referring compared to baseline Sa2VA (Yuan et al., 2025), and successfully captures both actor and target objects in interaction-aware referring setting.

6 Conclusion

We present InterRVOS, a new task that extends referring video object segmentation to explicitly model inter-object interactions by requiring segmentation of both actor and target entities described in paired expressions. To support this task, we introduce InterRVOS-8K, the first large-scale, automatically annotated dataset containing diverse and challenging interaction-aware expressions. We also propose a actor-target aware evaluation setting and a baseline architecture ReVIOSa that jointly segments both roles from a single expression. Our experiments demonstrate that modeling interactions leads to improved performance on complex video understanding benchmarks, highlighting the importance of interaction-centric supervision in RVOS.

APPENDIX

In appendix, we provide additional details and analyses that support the findings of the main paper. Section A outlines the implementation details, including model configurations and training settings, and models used in data annotation pipeline. Section B shows additional examples of InterRVOS-8K in Figure A1 and Figure A2. Section C presents comprehensive dataset statistics, emphasizing the linguistic and visual diversity of the InterRVOS-8K dataset. In Section D, we report additional experimental results that highlight the significance and effectiveness of our dataset and architecture, particularly in zero-shot generalization settings. Section E provides qualitative examples that demonstrate the model's ability to handle complex referring scenarios in videos. Section F describes the video clip extraction process used to prepare raw videos for annotation. Additionally, we provide details of the prompt used during data annotation process from Figure A7 to Figure A13, ensuring transparency in the data annotation pipeline. Finally, Sections G and H discuss the limitations of our work and its broader societal impact, including potential future directions.

A IMPLEMENTATION DETAILS

A.1 MODEL CONFIGURATIONS AND TRAINING SETTINGS

For the proposed architecture ReVIOSa, we utilize InternVL-2.5 (Chen et al., 2024) as the base model for multimodal large language model (MLLM), applying LoRA (Hu et al., 2022) tuning exclusively to enable the generation of the specialized tokens `[SEG_ACT]` and `[SEG_TAR]`. For the segmentation module, we adopt SAM2 (Ravi et al., 2024) and fine-tune only its decoder while keeping the image encoder frozen. The model is trained for 10 epochs with a batch size of 2. We report results using two model scales: InternVL-2.5-1B and InternVL-2.5-4B. The 1B model is trained on 4 NVIDIA RTX 3090 GPUs for 4 hours, whereas the 4B model is trained on 4 NVIDIA A6000 GPUs for 10 hours.

A.2 DATA ANNOTATION PIPELINE

In Section 3.1 of the main paper, we describe our data annotation pipeline as a four-stage process. Among these, **Stage 1** and **Stage 3** utilize GPT-4o (Hurst et al., 2024) 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 (Grattafiori et al., 2024). GPT ensures high-quality semantic understanding and detecting interactions, while the LLaMA model provides efficient and flexible generation of structured referring expressions. As a result, we are able to generate diverse, high-quality annotations at scale, while maintaining consistency across different stages of the annotation process.

B ADDITIONAL EXAMPLES OF INTERRVOS-8K

Figure A1 and Figure A2 present additional examples from the InterRVOS-8K 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 ADDITIONAL DATASET STATISTICS

The overall statistics of the InterRVOS-8K dataset are shown in Figure A3, capturing both linguistic and visual aspects of the data. 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 sufficient 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 detailed video clip extraction procedure is described in Section F.

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-8K dataset is well-suited for advancing research in referring video object segmentation and interaction-centric video understanding.

Furthermore, (f), (g), (h) provides an overall interaction-focused statistics within InterRVOS-8K. 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. Overall statistics of the InterRVOS-8K dataset can be found in Appendix C.

D Additional experimental results

In this section, we provide further experimental results supporting the impact of our dataset and architecture, with a particular focus on their generalization capabilities in zero-shot settings.

D.1 Performance analysis by caption type

The experiment closely follows the setting of Table 4 in the main paper, with one key distinction: we categorize the MeViS (Ding et al., 2023) dataset into five caption types using GPT-4o (Hurst et al., 2024), aligning with the classification scheme adopted in our proposed dataset. Specifically, the five categories are: (1) Single object expression (Appearance and Motion), (2) Single object expression (Appearance-only), (3) Single object expression (Motion-only), (4) Multi-object expression, and (5) Interaction expression. This setup allows for a more precise evaluation of the effectiveness of both the dataset and the model architecture. In Table A1, (a) compares the effect of different training datasets under a shared architecture, Sa2VA-1B (Yuan et al., 2025). Our dataset not only leads to strong performance on our own evaluation set but also generalizes well to MeViS, outperforming other comparison datasets. (b) presents results using our proposed architecture, ReVIOSa-1B. These results demonstrate the advantage of our subject-object-aware design, which enables more effective video understanding compared to existing models, particularly when dealing with interaction-centric captions.

D.2 Zero-shot evaluation

Table A2 presents zero-shot evaluation results of models trained on different datasets—ReVOS (Yan et al., 2024), Ref-SAV (Yuan et al., 2025), and InterRVOS-5K—on three standard RVOS benchmarks: MeViS (Ding et al., 2023), Ref-Youtube-VOS (Seo et al., 2020), and Ref-DAVIS (Khoreva et al., 2019). These results illustrate how much transferable video understanding each training dataset provides. The baseline model used in the comparisons is Sa2VA (Yuan et al., 2025).

Notably, the model trained on InterRVOS-5K, a subset of our full dataset, achieves the highest performance across all benchmarks, demonstrating the strong generalization capability of our interaction-rich data. Although these benchmarks primarily feature isolated object descriptions and lack explicit interaction cues, InterRVOS-5K still facilitates the learning of robust visual-language alignment. We also report the results of our proposed architecture, ReVIOSa-1B, trained on the same InterRVOS-5K data. Despite being specifically designed for interaction-aware segmentation, ReVIOSa-1B per-

Training Dataset		Architecture	Single (Appear. + Motion)	Single (Appear.)	Single (Motion)	Multi	Interaction	Overall
(a)	ReVOS	Baseline	52.1	37.6	41.0	**65.0**	51.4	49.1
	Ref-SAV		50.2	<u>39.5</u>	34.4	45.7	36.8	40.1
	InterRVOS-5K		<u>54.7</u>	37.9	<u>42.0</u>	62.3	<u>51.8</u>	<u>50.5</u>
(b)	InterRVOS-5K	ReVIOSa	**61.6**	**46.7**	**44.3**	<u>63.0</u>	**53.1**	**52.7**

Table A1: **Effectiveness of InterRVOS-8K and architecture.** We evaluate performance across five caption types. The models are trained on three different datasets, ReVOS (Yan et al., 2024), Ref-SAV (Yuan et al., 2025) and InterRVOS-5K, and evaluated on MeViS (Ding et al., 2023) *valid_u* set. (a) presents results using the baseline model, Sa2VA (Yuan et al., 2025), while (b) presents results using the ReVIOSa. Best and second-best results are shown in **bold** and <u>underline</u>, respectively. "Appear." denotes appearance.

Setting	Baseline			ReVIOSa
	ReVOS (Yan et al., 2024) (Videos 600 / Exps. 29K)	Ref-SAV (Yuan et al., 2025) (Videos 37k / Exps. 72K)	InterRVOS-5K (Videos 5K / Exps. 71K)	InterRVOS-5K (Videos 5K / Exps. 71K)
MeViS valid	39.6	32.8	**41.8**	<u>41.5</u>
MeViS valid_u	49.1	40.1	<u>50.5</u>	**52.7**
Ref-Youtube-VOS	57.5	54.2	**61.8**	<u>59.3</u>
Ref-DAVIS	62.8	62.1	**67.1**	<u>63.7</u>

Table A2: **Zero-shot evaluation on standard RVOS benchmarks.** This table compares the generalization ability of models trained on three datasets by evaluating them in a zero-shot manner on conventional RVOS benchmarks: MeViS (Ding et al., 2023), Ref-Youtube-VOS (Seo et al., 2020), and Ref-DAVIS (Khoreva et al., 2019). The baseline model is Sa2VA (Yuan et al., 2025), and InterRVOS-5K consistently outperforms models trained on other datasets, demonstrating the effectiveness of our interaction-centric data. We also report results from ReVIOSa-1B trained on InterRVOS-5K, which show that even without specific adaptation to interaction-sparse benchmarks, the model maintains competitive performance.

forms competitively even on interaction-scarce benchmarks, highlighting the generalizability of our framework.

E QUALITATIVE RESULTS

We present qualitative results to demonstrate the effectiveness of our proposed model in handling complex, interaction-centric referring expressions. Figure A4 compares our model (ReVIOSa) with a strong baseline (Sa2VA) on the proposed InterRVOS-8K 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 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 perform joint subject-object inference within a single forward pass. As illustrated in Figure A5 and A6, the model utilizes dedicated `[SEG_ACT]` and `[SEG_TAR]` tokens to simultaneously localize both the subject and the object described in interaction-centric expressions. This dual segmentation capability enables our model to effectively capture relational semantics and dynamic interactions between entities. Such ability opens up opportunities for downstream applications such as human-object interaction understanding, social behavior analysis, and fine-grained activity reasoning in videos.

These qualitative results collectively validate the robustness, flexibility, and extensibility of our approach in real-world video understanding tasks that require precise multi-entity segmenting guided by natural language.

F VIDEO CLIP EXTRACTION PROCEDURE

The InterRVOS-8K dataset is constructed using source videos from the VidOR dataset (Shang et al., 2019), which contains a large number of long-form videos, many exceeding 1,000 frames in length.

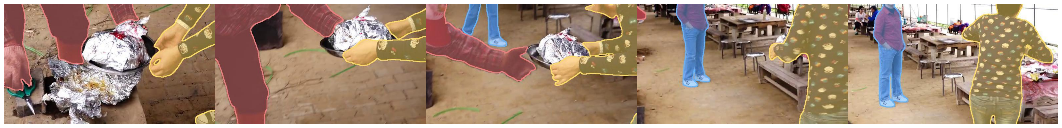
To generate more diverse and effective video clips for referring video object segmentation, we apply a systematic clip extraction strategy. Specifically, each original source video is divided into non-overlapping temporal bins of 1,000 frames. From these, we select only the first and last bins to increase 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 scene 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.

G LIMITATIONS

Despite the scalability and effectiveness of our data annotation pipeline, certain challenges remain. Obtaining source videos that are both dynamic and rich in object interactions can be difficult, which may limit the diversity of interaction scenarios. Additionally, in cases where object boundaries are unclear, such as under motion blur or corrupted datas, the quality of segmentation or generated descriptions may be affected. Nonetheless, these challenges are common in real-world video data and highlight opportunities for future improvement.

H BROADER IMPACT

InterRVOS introduces a large-scale, interaction-aware dataset and a action-target segmentation framework that together enable fine-grained understanding of visual relationships. The dataset provides structured referring expressions paired with distinct masks for both the subject and object in an interaction, while the architecture explicitly models these roles through dedicated segmentation tokens. This synergy allows models to interpret complex multi-entity instructions such as "the person handing a tool to the worker", fostering more intuitive and reliable human-AI communication.



[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]

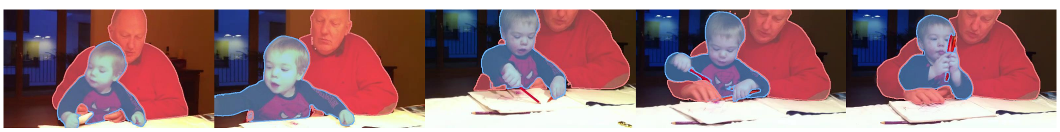
Actor [0] / Target [1]

"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 A1: **Examples of InterRVOS-8K.**

[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"

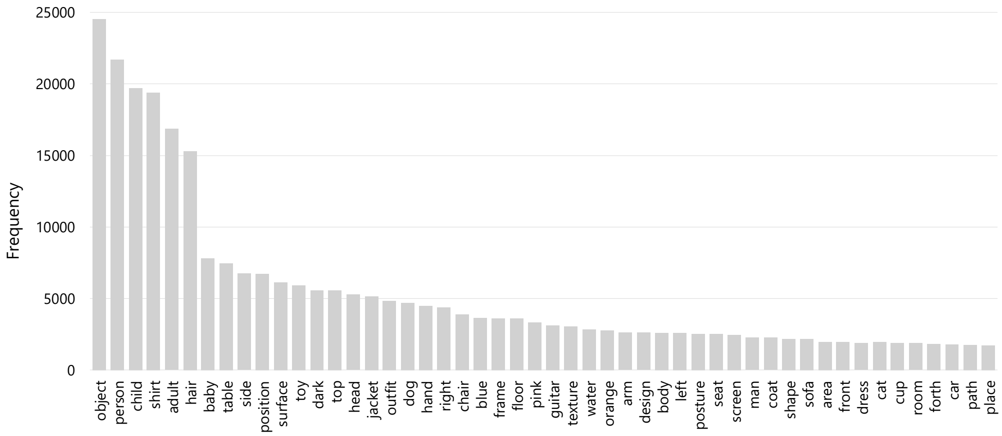
"Adult in white shirt explaining to adult in light-colored shirt and sporty white car"

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

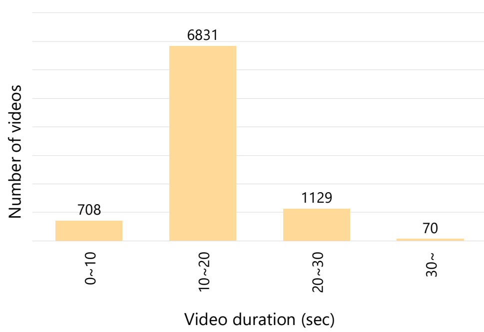
"Car being discussed by people"

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

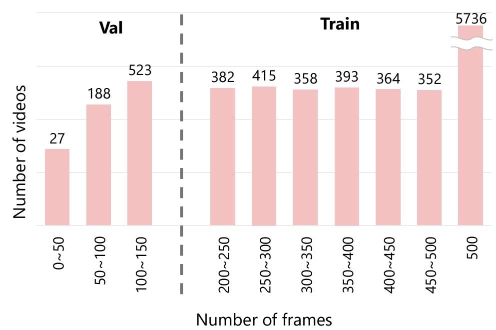
Figure A2: **Examples of InterRVOS-8K.**



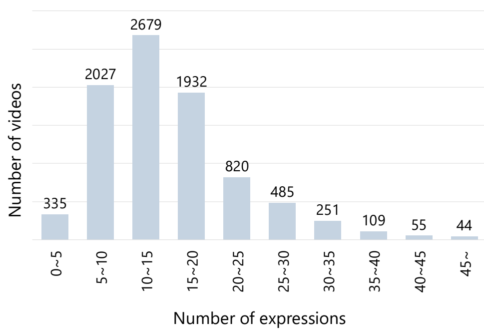
(a) Frequency distribution of the Top-50 most frequent words



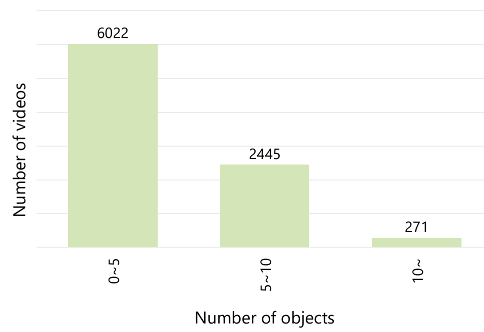
(b) Distribution of video duration



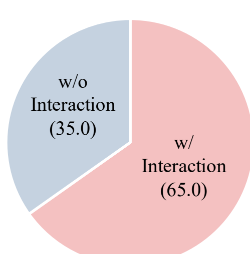
(c) Distribution of video frames



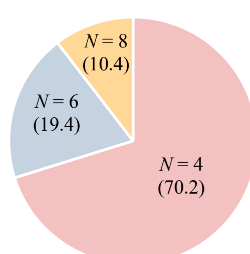
(d) Number of expressions per video



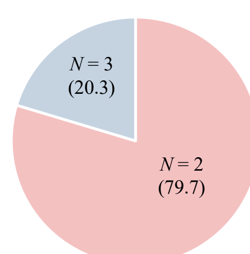
(e) Number of objects per video



(f) Videos with interaction expressions



(g) Interaction expressions per video



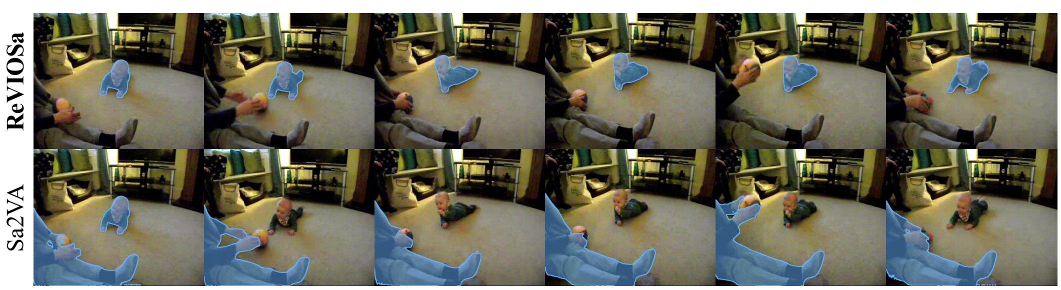
(h) Objects engaged in one interaction

Figure A3: **Overall statistics of InterRVOS-8K.**

"Hand with a black brush brushing a gray tabby cat"

"The young child with short, wet hair sitting in the bathtub, occasionally adjusting their posture slightly"

"Pink bottle being used by child in yellow shirt"

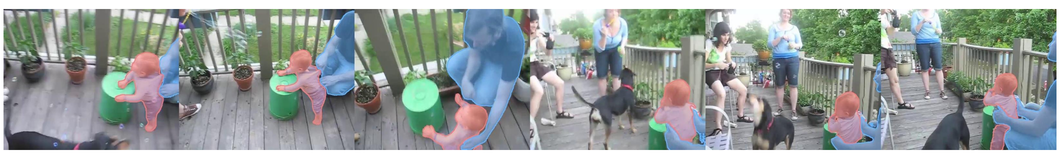


"The small fair-skinned baby in a dark outfit and patterned bib crawling forward on the carpeted floor, shifting its gaze and moving slightly to the left"

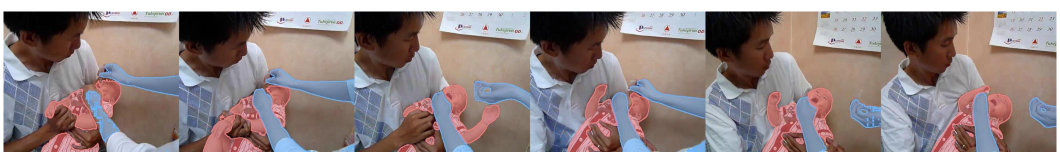
Figure A4: **Qualitative results.** Qualitative comparisons between our model (InterRVOS) and the baseline model (Sa2VA) on the proposed InterRVOS-8K dataset for the RVOS task. InterRVOS consistently produces more accurate and temporally consistent segmentations, especially in challenging scenarios involving fine-grained interactions, appearance ambiguity, or motion. These results demonstrate the effectiveness of InterRVOS in aligning linguistic cues with visual targets across time.

User: Can you segment the adult in patterned swim trunks guiding child in dark blue shorts? If both an actor and target objects are present, please provide separate segmentation masks for each.
ReVIOSa: It is [SEG_ACT] and [SEG_TAR].

User: Please segment furry dog pushing colorful toy with green handle. If both an actor and target objects are present, return individual segmentation masks for the subject and the object.
ReVIOSa: Sure, [SEG_ACT] and [SEG_TAR].



User: What is man with short hair holding a baby with light-colored hair? Please respond with segmentation mask. If both an actor and target objects are present, ensure that each is segmented independently.
ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

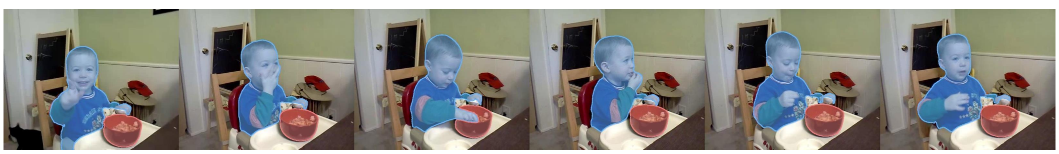


User: What is bare-skinned hand feeding the baby in red and white stripes? Please output segmentation mask. If both an actor and target objects are present, please segment them separately and output both masks.
ReVIOSa: Sure, the segmentation result is [SEG_ACT] and [SEG_TAR].

Figure A5: **Qualitative results.** Joint actor-target segmentation results using our proposed model with interaction-centric referring expressions on the InterRVOS-8K dataset. Leveraging dedicated `[SEG_ACT]` and `[SEG_TAR]` tokens, our model is able to segment both the actor (blue) and the target (pink) entities within a single forward pass. Each example corresponds to a complex expression describing an interaction between two entities. These results demonstrate the model's ability to localize and distinguish multiple semantically linked objects simultaneously, showing potential for downstream applications such as human-object interaction understanding, social activity recognition, and fine-grained video scene interpretation.

User: Please identify and segment the adult in dark jacket guiding child in helmet. If both an actor and target objects are present, please distinguish between them and generate separate masks accordingly.
ReVIOSa: [SEG_ACT] and [SEG_TAR].

User: Can you highlight the blonde child in a pink and red dress reaching towards a small black dog? If both an actor and target objects are present, generate and return individual masks for each of them.
ReVIOSa: It is [SEG_ACT] and [SEG_TAR].



User: Where is the young child with colorful shirt eating from red bowl? If both an actor and target objects are present, please provide separate segmentation masks for each.
ReVIOSa: Sure, it is [SEG_ACT] and [SEG_TAR].

User: Please segment child in dark blue shorts being guided by adult in patterned swim trunks. If both an actor and target objects are present, return individual segmentation masks for the subject and the object.
ReVIOSa: Sure, the segmentation result is [SEG_ACT] and [SEG_TAR].

Figure A6: **Qualitative results.** Joint actor-target segmentation results using our proposed model with interaction-centric referring expressions on the InterRVOS-8K dataset. Leveraging dedicated `[SEG_ACT]` and `[SEG_TAR]` tokens, our model is able to segment both the actor (blue) and the target (pink) entities within a single forward pass. Each example corresponds to a complex expression describing an interaction between two entities. These results demonstrate the model's ability to localize and distinguish multiple semantically linked objects simultaneously, showing potential for downstream applications such as human-object interaction understanding, social activity recognition, and fine-grained video scene interpretation.

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 A7: **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 A8: **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.

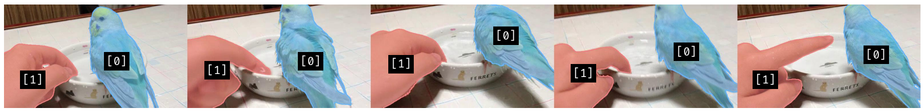
```
)

"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 A9: **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)

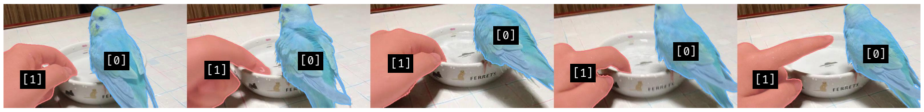


<task>
You are given a video in which multiple labeled objects appear. Your task is to identify any visible interaction between the labeled objects, determine the type and direction of interaction, and describe it appropriately.
</task>
<objectives>
1. Determine whether any interaction is visually observable 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 A10: **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 A11: **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 A12: **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 A13: **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

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.

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

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.

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 (ICCV)*, pp. 16321–16330, October 2021.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2694–2703, 2023.

Qi Fan, Yanghao Liu, Wei Wang, Ning Xu, Dahua Lin, Alan Yuille, and Chen Change Loy. Moma: A multi-object multi-action dataset for understanding human activities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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*, pp. 5958–5966, 2018.

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 e-prints*, pp. arXiv–2407, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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.

Mengmeng Ji, Jingwei Wang, Xing Xu, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Action genome: Actions as compositions of spatiotemporal scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10236–10247, 2020.

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*, pp. 123–141. Springer, 2019.

Hao Liu, Jiapeng Yang, Mingyu Wang, and et al. Videoglamm: Video grounded language-image pretraining with masked modeling. *arXiv preprint arXiv:2411.04923*, 2024.

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

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*, pp. 920–930, 2023.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Refer-youtube-vos: A dataset for video object segmentation with language referring expressions. In *ECCV*, 2020.

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.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pp. 208–223. Springer, 2020.

Xindi Shang, Linchao Li, Yao Zhang, Tao Jiang, Xin Yang, and Zhiwu Chen. Video relationship detection. In *ACM International Conference on Multimedia (ACM MM)*, pp. 1074–1082, 2017.

Xindi Shang, Xiyang Ren, Linchao Li, Zhiwu Chen, Yu-Gang Liu, and Yu Zhou. Annotating objects and relations in user-generated videos. In *ACM International Conference on Multimedia (ACM MM)*, pp. 1308–1316, 2019.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Jianhao Wang, Feng Gao, Jianke Zhu, and Qionghai Dai. Soc: Segmenting objects by categories for referring video object segmentation. *arXiv preprint arXiv:2305.17011*, 2023.

Renrui Wang, Ze Ma, Xiang Li, and et al. Villa: Unifying vision-language segmentation tasks with large multi-modal models. *arXiv preprint arXiv:2407.14500*, 2024.

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

Yifei Wu, Guojun Yu, Wenqi Jia, Qi Jin, Jie Zhou, and Yu Qiao. Star: Structured action understanding in instructional videos. *arXiv preprint arXiv:2405.09711*, 2024.

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

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

Jianyuan Zhou, Jing Gao, Hui Zeng, and Huchuan Lu. Dshmp: Densely supervised hierarchical mask propagation for referring video object segmentation. *arXiv preprint arXiv:2404.03645*, 2024.