ReVSeg: Incentivizing the Reasoning Chain for Video Segmentation with Reinforcement Learning

Yifan Li[1,2] Yingda Yin[3*†] Lingting Zhu[3†] Weikai Chen[3] Shengju Qian[3]
Xin Wang[3] Yanwei Fu[1,2*]
[1]Fudan University [2]Shanghai Innovation Institute [3]LIGHTSPEED

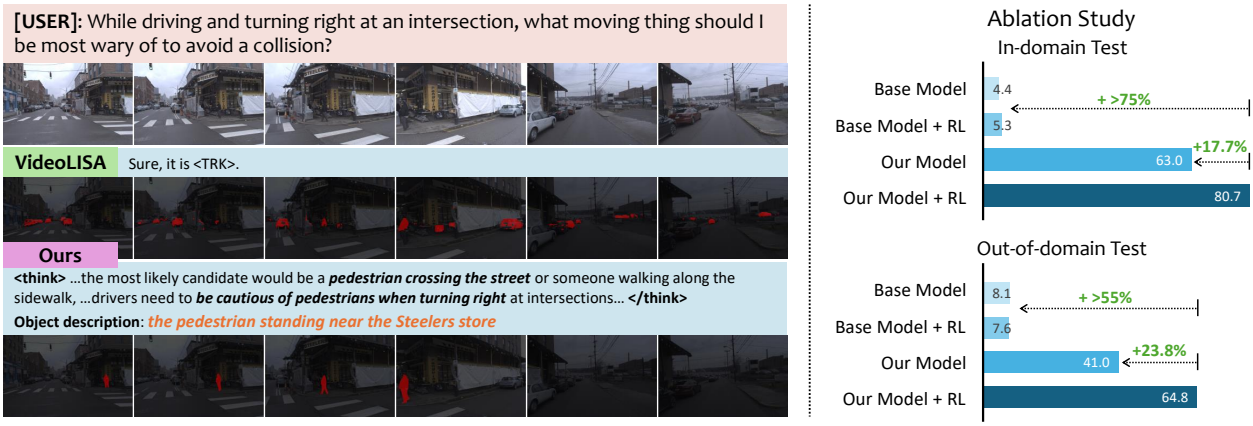

Figure 1. (Left) Through an explicit reasoning chain, our ReVSeg tackles reasoning-focused video object segmentation and accurately grounds objects referenced by complex, abstract real-world queries. (Right) While the base model and its RL variant struggle on the task, our method achieves strong performance, with RL post-training yielding a further substantial boost. We report the $\mathcal{J}\&\mathcal{F}$ metric on Ref-DAVIS17 (in-domain) and ReasonVOS (out-of-domain) datasets in the chart.

Abstract

Reasoning-centric video object segmentation is an inherently complex task: the query often refers to dynamics, causality, and temporal interactions, rather than static appearances. Yet existing solutions generally collapse these factors into simplified reasoning with latent embeddings, rendering the reasoning chain opaque and essentially intractable. We therefore adopt an explicit decomposition perspective and introduce ReVSeg, which executes reasoning as sequential decisions in the native interface of pretrained vision language models (VLMs). Rather than folding all reasoning into a single-step prediction, ReVSeg executes three explicit operations – semantics interpretation, temporal evidence selection, and spatial grounding – aligning pretrained capabilities. We further employ reinforcement learning to optimize the multi-step reasoning chain, enabling the model to self-refine its decision quality from outcome-driven signals. Experimental results demonstrate that ReVSeg attains state-of-the-art performances on standard video object segmentation benchmarks and yields interpretable reasoning trajectories. Project Page.

* Corresponding authors.
† Project leads.

1. Introduction

Human interpretation of videos relies on recognizing how events unfold, why actions occur, and when key moments matter. Traditional video object segmentation (VOS) methods, however, primarily exploit appearance-only or category-level cues [3, 29, 40, 45, 46]. Reasoning VOS elevates the task: the model must parse abstract, underspecified instructions, draw on commonsense and causal reasoning, integrate temporal dynamics with semantic knowledge to identify the target (e.g., "the runner most likely to win" or "the object causing the accident").

While recent vision-language models (VLMs) have shown promising capabilities for reasoning segmentation [2, 16, 50], current systems predominantly model reasoning VOS as a single-step latent prediction: inserting a special token (e.g., <SEG>) and decoding it directly into mask outputs [2, 8, 20, 35, 41, 42, 50, 53]. This collapses the multi-step reasoning process — interpreting abstract instructions, identifying candidates, and spatial-temporal

grounding — into a simple conclusion as well as opaque embeddings. Such compactness comes at a great cost: interpretability is bounded, distribution shift arises from forcing VLMs into non-native output spaces, and substantial data is required for supervised fine-tuning.

These challenges reflect a deep insight: effective video reasoning unfolds through *a sequence of deliberate choices*, not a single latent inference. A model must determine *where* to focus, *when* to attend, and *which* entity the query refers to. Motivated by this principle, we introduce ReVSeg, which eliminates the use of latent segmentation tokens and instead reformulates reasoning video segmentation as an explicit sequence of reasoning chain. In particular, ReVSeg decomposes reasoning VOS into three actions aligned with VLM-native capabilities: video understanding (interpreting the query and assessing scene dynamics), temporal grounding (identifying key frames or intervals pertinent to the query), and spatial grounding (localizing the target objects within selected frames). ReVSeg orchestrates these primitives through multi-turn actions with a single VLM, ensuring the semantic context established in early reasoning steps seamlessly propagates to downstream localization. Executing reasoning through these native interfaces preserves pretraining alignment, avoids heavy re-training, and transforms an otherwise entangled problem into a structured procedure.

Once the reasoning chain is made explicit, the central challenge becomes how to optimize the chain itself. Supervised learning provides little leverage, as the correctness of intermediate decisions is hardly annotated. We therefore adopt *reinforcement learning* (RL) [9, 30], viewing reasoning VOS as a behavioral policy that should be rewarded only when its full decision trajectory leads to a correct segmentation outcome [17]. Yet, naïvely applying RL encounters obstacles: reasoning VOS offers extremely sparse success signals, and requires coordination across video understanding, temporal selection, and spatial localization. To address these challenges, we introduce reasoning-aligned rewards which inject signals at critical decision points, ensuring that RL enhances the model's reasoning behavior itself. Together with decomposition, RL transforms the task from an opaque latent regression into a structured, optimizable reasoning policy: decomposition isolates what the model must decide while RL determines how those decisions should cooperate. This synergy yields a reasoning policy that is both internally consistent and empirically strong.

Experiments show that ReVSeg delivers the state-of-the-art (SOTA) performances on multiple standard VOS benchmarks, outperforming both latent-based VLMs [8, 16, 20, 35, 41, 42, 50, 53, 55] and the training-free explicit VLM [13], with auditable reasoning traces. As shown in Figure 1, controlled ablations demonstrate that without our decomposed reasoning chain, both the base VLM and its RL post-trained variant tend to fail on the complex reasoning VOS task. In contrast, our framework attains strong performance even before RL, and the RL post-training further delivers a substantial improvement. In summary, our contributions are as follows:

- *A principled decomposition of reasoning VOS.* We introduce ReVSeg, which **re**formulates **re**asoning **v**ideo **seg**mentation as an explicit multi-step reasoning chain built from native VLM primitives.
- *An RL framework that optimizes the reasoning chain itself.* We develop a novel RL-based formulation that directly optimizes the reasoning chain, allowing the model to refine decision quality without requiring dense supervision.
- We set the new state of the art on standard VOS benchmarks while providing interpretable reasoning traces.

2. Related Works

2.1. Reasoning Video Object Segmentation

Fine-grained text-guided VOS has advanced through specialized segmentation architectures and referring-video pipelines [3, 5, 29, 45]. For the emerging and more complex setting of reasoning VOS, recent methods fine-tune VLMs to emit implicit mask tokens [2, 8, 20, 41, 42, 50, 53], often paired with strong decoders such as SAM/SAM2 [15, 26]. Yet robust reasoning in complex videos—multiple similar objects, occlusion, fast motion, and long-range context—remains challenging. Explicit textual reasoning via CoT has begun to appear, *e.g.*, the training-free framework CoT-RVS [13], which adopts two independent VLM systems for frame-by-frame segmentation bridged via text. While effective, such designs are bounded by the separated information flow and module interoperability and is constrained to be training-free. In this work, we formulate reasoning VOS as an explicit reasoning chain within a unified VLM, and employ reinforcement learning for self-improvement.

2.2. VLMs with Reinforcement Learning

Reinforcement learning has emerged as a practical route to strengthen the reasoning capabilities of large models when supervised annotated data are scarce. Test-time scaling via Chain-of-Thought [43] improves VLM reasoning, while rewards further refine solution quality. Notably, Group Relative Policy Optimization (GRPO) [30] enables efficient critic-free updates and strong gains with modest training budgets, as demonstrated by DeepSeek-Zero [9]. Recent efforts [11, 24, 32, 37, 47] leverage VLMs for high-level reasoning, and reasoning image grounding [21, 22, 27, 48, 54] has been actively studied. However, in the video modality, most work [4, 7, 17] still targets high-level understanding; fine-grained spatio-temporal grounding remains limited due to task complexity. Our results show that reinforcement

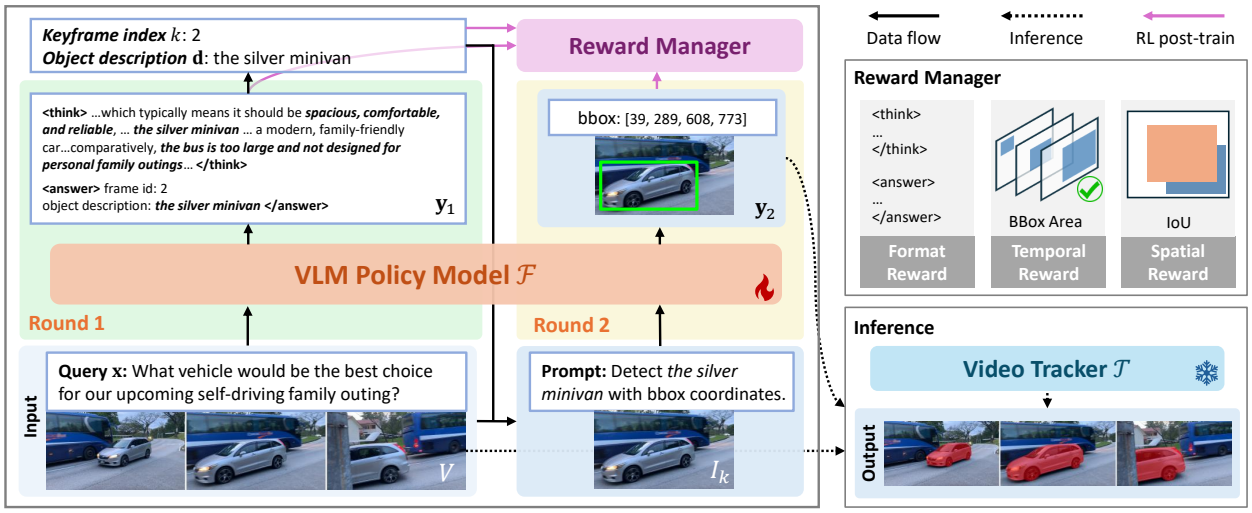

Figure 2. Overview of ReVSeg. The model runs a two-turn reasoning chain over the input video and query. Round one analyzes the scene and selects an informative keyframe with a concise object description. Round two grounds the target on that keyframe by predicting a bounding box. The keyframe-bbox pair conditions a video tracker to produce full segmentation sequence. A reward manager provides concise signals to post-train the VLM via reinforcement learning, improving keyframe selection, grounding accuracy, and overall robustness.

learning can effectively boost reasoning for video grounding.

3. Method

3.1. Overview

Task Formulation. Given a natural language query $\mathbf{x}$ and a video sequence with T frames $V = \{I_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$, where H and W denote the height and width of each frame, the goal of reasoning VOS is to segment the query-referred objects throughout the video. The model predicts a sequence of binary masks $M = \{m_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W}$, where $m_t \in \{0,1\}^{H\times W}$ represents the foreground region in frame t. Let $M^* = \{m_t^*\}_{t=1}^{T}$ denote the ground-truth masks. The task is to learn a mapping

$$(V, \mathbf{x}) \rightarrow M, \tag{1}$$

that maximizes agreement with M^*. This mapping must correctly resolve the semantic correspondence specified in query and maintain temporal consistency across the video.

Pipeline Overview. ReVSeg formulates reasoning-centric video object segmentation as an explicit sequence of reasoning steps. The complex task is decomposed into three VLM-native capabilities – video understanding, temporal grounding, and spatial grounding – and executed through multi-turn dialogue with a single VLM.

As illustrated in Fig. 2, the **first-round dialogue** takes the video and query $(V, \mathbf{x})$ as input. The VLM performs video understanding and temporal grounding: it interprets the abstract query, analyzes scene dynamics, produces a concise spatial description of the target, and identifies the keyframe that best captures the entity of interest. Within this round, the raw, cluttered video sequence and the vague high-level query (e.g., "the best choice for the family outing") are distilled into a well-specified keyframe and a concrete textual description (e.g., "the silver minivan"). These intermediate outputs encapsulate the VLM's commonsense and causal reasoning, effectively transforming a complex video–text reasoning problem into a substantially simpler image-level segmentation task.

Given these intermediate results, the **second-round dialogue** processes the selected keyframe and the concrete object description. The VLM then performs spatial grounding and outputs a tight bounding box for the specified object, completing the two-round reasoning chain. With the localized keyframe target, an off-the-shelf video tracker (*e.g.*, SAM2) is readily applied to produce the final video mask predictions.

Because both rounds occur within a single VLM, the semantic context established during early reasoning is seamlessly propagated to subsequent steps, ensuring consistency throughout the chain.

Finally, we optimize the reasoning process through reinforcement learning. Rather than rewarding only the final outcome of the complex task, the decomposition of the pipeline and the availability of meaningful intermediate products allow for richer, reasoning-aligned reward signals that better guide policy improvement.

3.2. Decomposed Generation with Reasoning Chain

Directly producing spatio-temporal grounding from video input remains challenging for current VLMs [1, 11, 31, 32, 36, 38]. Owing to the limited availability of high-quality annotated datasets, these models have not been extensively

pretrained for VOS. Even with test-time scaling strategies – such as Chain-of-Thought prompting [43], multi-rollout sampling with self-consistency [39], or search-based inference methods including Tree Search and Beam Search [33, 51] – VLMs still struggle to generate reliable spatio-temporal grounding results directly from raw video.

These limitations highlight a key observation that complex video reasoning inherently unfolds as a sequence of interdependent reasoning actions, rather than a single-step prediction. This insight motivates us to decompose the reasoning-based VOS task into a set of primitive capabilities, orchestrated through a multi-turn reasoning chain, as described below.

First Round Rollout. In the first reasoning round, the VLM $\mathcal{F}$ receives a video–query pair $(V, \mathbf{x})$ and generates a text response $\mathbf{y_1}$ under an instruction-guided prompt. During this step, the model interprets the user query, analyzes the video content, infers the target entity, identifies its temporal occurrence within the sequence, and produces a concise textual description of the target object grounded in a keyframe. The output of this reasoning round can be formalized as:

$$\mathbf{y_1} \sim \mathcal{F}(\cdot \mid V, \mathbf{x}). \tag{2}$$

A parser $\mathcal{G}$ then processes the structured response $\mathbf{y_1}$, extracting: the keyframe index $k \in \{0, 1, \cdots, T-1\}$, a concise spatial description $\mathbf{d}$ of target objects, and a status flag $S_1 \in \{\texttt{succ}, \texttt{fail}\}$ indicating the extraction success:

$$\mathcal{G}(\mathbf{y_1}) = \begin{cases} (S_1, k, \mathbf{d}), & S_1 = \texttt{succ} \\ (S_1, \texttt{null}, \texttt{null}), & S_1 = \texttt{fail}. \end{cases} \tag{3}$$

Based on the selected keyframe index k, the corresponding frame I_k is retrieved from the video sequence V.

Second Round Rollout. If the status flag from the previous round is $S_1 = \texttt{succ}$, the rollout generation proceeds to the second stage. In this round, the VLM $\mathcal{F}$ receives the selected keyframe I_k and the concise object description $\mathbf{d}$, together with the generation history $(V, \mathbf{x}, \mathbf{y_1})$ from the first round. Conditioned on this context and an instruction-guided prompt, the model generates a text response $\mathbf{y_2}$. At this stage, $\mathcal{F}$ is responsible for spatial grounding: it localizes the target object in keyframe I_k using both the visual input and the accumulated dialogue history. The output $\mathbf{y_2}$ follows a structured format, formalized as:

$$\mathbf{y_2} \sim \mathcal{F}(\cdot | V, \mathbf{x}, \mathbf{y_1}, I_k, \mathbf{d}). \tag{4}$$

Similarly, the parser $\mathcal{G}$ processes the structured response $\mathbf{y_2}$, extracting a bounding box $B_k \in \mathbb{R}^4_{\geq 0}$ and a new status flag S_2:

$$\mathcal{G}(\mathbf{y_2}) = \begin{cases} (S_2, B_k), & S_2 = \texttt{succ} \\ (S_2, \texttt{null}), & S_2 = \texttt{fail}. \end{cases} \tag{5}$$

The final rollout output o is obtained by concatenating the two-stage responses:

$$o = \mathbf{y_1} \oplus \mathbf{y_2}. \tag{6}$$

This two-round decomposition cleanly separates video understanding and temporal selection from spatial localization, facilitates well-defined task at each stage, and provides a stable, modular interface for the RL optimization.

3.3. Reasoning VOS with GRPO

Reasoning capability is the key factor determining the upper bound of reasoning-based VOS. However, the supervised reasoning data required by Supervised Fine-tuning (SFT) is scarce and costly. Inspired by outcome-driven training in recent reasoning systems [9, 30], we employ reinforcement learning to enable the policy model $\mathcal{F}$ to self-improve under task-specific rewards, thereby enhancing its reasoning ability.

Reward Modeling. Rewards shape the optimization dynamics and are therefore crucial in RL. Following the minimalist design philosophy of DeepSeek-R1-Zero [9], we adopt a rule-based reward system consisting of three components:

- *Format Reward* r_f: Correct output formatting is essential for interacting with the environment. The model must place its reasoning process between `<think>` and `</think>` tags and the final answer between `<answer>` and `</answer>` tags. In addition, the first-turn output $\mathbf{y_1}$ must include the keyframe index k and object description $\mathbf{d}$ in JSON format, while the second-turn output $\mathbf{y_2}$ must provide the bounding box B_k in JSON. Based on the degree to which the output o satisfies these rules, the format reward r_f is assigned a value in $[0, 1]$.
- *Temporal Reward* r_t: The keyframe I_k selected by $\mathcal{F}$ in $\mathbf{y_1}$ critically affects subsequent spatial grounding. Beyond merely containing the target object, we encourage selecting frames where the object is clearly visible, minimally occluded, and sufficiently large. We experiment with several temporal reward choices and finally choose the normalized area of the ground-truth bounding box at I_k:

$$r_t = \mathbb{1}_{(||m_k^*||_1 > 0)} \cdot \frac{\mathcal{S}(m_k^*) - \min_t \mathcal{S}(m_t^*)}{\max_t \mathcal{S}(m_t^*) - \min_t \mathcal{S}(m_t^*)}. \tag{7}$$

  where $\mathcal{S}(m_t^*)$ denotes the pixel area of the ground truth bbox at frame I_t and $\mathbb{1}_{(\cdot)}$ denotes the indicator function. See Sec. 4.3 for ablations.
- *Spatial Reward* r_s: This reward measures final detection quality of the predicted B_k. Following prior works [21, 22], we use Intersection-over-Union (IoU) between the predicted and ground-truth boxes. If IoU > 0.5, the prediction is considered correct and $r_s = 1$; otherwise $r_s = 0$.

Table 1. Reasoning video object segmentation performance comparison on ReasonVOS [2] dataset.

Method		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
MTTR [3]	[CVPR'22]	29.1	33.1	31.1
ReferFormer [46]	[CVPR'22]	30.2	35.6	32.9
SOC [23]	[NeurIPS'24]	33.3	38.5	35.9
OnlineRefer [44]	[CVPR'23]	34.6	42.9	38.7
SgMg [25]	[ICCV'23]	33.7	38.7	36.2
LISA [16]	[CVPR'24]	29.1	33.1	31.1
VideoLISA [2]	[NeurIPS'24]	45.1	49.9	47.5
GLUS [20]	[CVPR'25]	47.5	52.4	49.9
RGA-3B [35]	[ICCV'25]	49.1	54.3	51.7
RGA-7B [35]	[ICCV'25]	51.3	56.0	53.6
CoT-RVS-online-7B [13]	[arXiv'25]	49.5	54.5	52.0
CoT-RVS-offline-13B [13]	[arXiv'25]	47.5	54.0	50.7
ReVSeg-7B (Ours)	-	**61.8**	**67.7**	**64.8**

The total reward for an output o combines these components with status flags S_1 and S_2 indicating whether each step succeeds:

$$r = r_f + \mathbb{1}_{(S_1=\text{succ})} r_t + \mathbb{1}_{(S_1=\text{succ } \& \ S_2=\text{succ})} r_s. \quad (8)$$

Objective. We adopt Group Relative Policy Optimization (GRPO) [30], a critic-free variant of Proximal Policy Optimization (PPO) [28] tailored for sequence models. Given an input query $q = \{V, \mathbf{x}\}$, GRPO samples a group of n candidate outputs $\{o_i\}_{i=1}^{n}$ from the current policy π_{old}, evaluates their rewards $\{r_i\}$, and computes a normalized within-group advantage:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_n)}{\text{std}(r_1, \ldots, r_n)}. \quad (9)$$

Since we use on-policy updates ($\pi_{\text{old}} = \pi_\theta$) in practice, importance sampling ratios are remains at one. The policy loss, combined with KL regularization to a reference model π_{ref}, yields the final training objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}\sim\pi_\theta(\cdot|q)} \left[\frac{1}{n} \sum_{i=1}^{n} \left(A_i - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right) \right]. \quad (10)$$

4. Experiments

4.1. Experiment Settings

Training Datasets. For our efficient RL post-training, we rely solely on the video object segmentation (VOS) data, in contrast to previous works [2, 8, 41, 42, 50] that jointly fine-tunes on large, heterogeneous corpora spanning video segmentation, image segmentation, and VQA datasets. Specifically, we curate training data from five benchmarks: Ref-YouTube-VOS [29], MeViS [5], Ref-DAVIS17 [14], ReVOS [50] and LV-VIS [34]. For each annotated sequence, we first convert per-frame masks to bounding boxes, which serve as ground-truth signals for post-training rewards. To ensure label quality, we run SAM2 [26] on every frame conditioned on its ground-truth box to obtain predicted masks, compute IoU against the annotated masks, and discard all queries from any video whose mean IoU falls below 0.6. This filtering yields approximately 67k data pairs.

Benchmarks. We evaluate on five standard VOS benchmarks: two reasoning datasets including ReVOS [50] and ReasonVOS [2] and three referring datasets including Ref-DAVIS17 [14], Ref-YouTube-VOS [29], MeViS [5]. Notably, ReasonVOS has no training split, thus its evaluation is zero-shot, providing a clearer measure of the model's generalization ability.

Baselines. We benchmark against three families of methods to contextualize our gains. (1) *Segmentation Specialists*: strong VOS/Ref-VOS systems [3, 5, 6, 10, 18, 23, 25, 29, 44, 46] trained with dense supervision, optimized for mask quality and temporal consistency. (2) *VLM-Based with Latent Tokens Methods*: methods [2, 8, 20, 35, 41, 42, 50, 53] that fine-tune VLMs to emit task-specific control tokens or logits that drive a downstream mask head, which are the current mainstream for reasoning VOS. (3) *VLM-Based with Explicit Reasoning Methods*: methods that perform reasoning to explicitly ground targets via boxes/masks, an under-explored paradigm where CoT-RVS [13] and our method fall. We report results across all three to isolate the advantage of our proposed ReVSeg.

Evaluation Metrics. Following previous works [2, 8, 13, 20, 41, 42, 50, 53] on reasoning video object segmentation, we report region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$) and their mean ($\mathcal{J}\&\mathcal{F}$) as the primary video-level metrics.

Implementation Details. We adopt Qwen2.5-VL-7B [1] as the default reasoning model $\mathcal{F}$ and SAM2 (Hiera-L) [26] as the default video tracker model $\mathcal{T}$. For post-training $\mathcal{F}$ with GRPO, each optimizer step processes 128 input data, and we sample $n = 8$ rollouts per prompt, yielding an effective batch of 1024 sequences per optimizer step. The learning rate is set to $1e-6$, and the KL regularization coefficient $\beta = 1e-3$. For each video, we uniformly sample 16 frames as input to $\mathcal{F}$. All input frames are resized to 448×448 before the first round generation. In the second round, the selected keyframe I_k is resized to 840×840 for spatial grounding. The tracker $\mathcal{T}$ operates on the full video at its original resolution.

4.2. Experimental Results

***Reasoning* Video Object Segmentation.** We first evaluate on reasoning VOS benchmarks – ReasonVOS dataset in Tab. 1 and ReVOS dataset in Tab. 2. On ReasonVOS, ReVSeg-7B achieves a decisive margin over the previous state-of-the-art (SOTA) method RGA-7B [35], improving $\mathcal{J}$ by +10.5 points, $\mathcal{F}$ by +11.7 points, and $\mathcal{J}\&\mathcal{F}$ by

Table 2. Reasoning video object segmentation performance comparison on ReVOS [50] dataset.

Method		Type	referring			reasoning			overall		
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
MTTR [3]	[CVPR'22]	Segmentation Specialists	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5
ReferFormer [46]	[CVPR'22]		31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1
LMPM [5]	[ICCV'23]		29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4
LLaMA-VID [18] + LMPM [5]	[ECCV'24]		29.0	39.1	34.1	12.8	23.7	18.2	20.9	31.4	26.1
LISA-7B [16]	[CVPR'24]	VLM-Based w/ Latent Tokens	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9
LISA-13B [16]	[CVPR'24]		45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6
TrackGPT(IT)-7B [55]	[arXiv'24]		46.7	49.7	48.2	36.8	41.2	39.0	41.8	45.5	43.6
TrackGPT(IT)-13B [55]	[arXiv'24]		48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0
VISA-7B [50]	[ECCV'24]		51.1	54.7	52.9	36.7	41.7	39.2	43.9	48.2	46.1
VISA-13B [50]	[ECCV'24]		52.3	55.8	54.1	38.3	43.5	40.9	45.3	49.7	47.5
VISA(IT)-7B [50]	[ECCV'24]		49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9
VISA(IT)-13B [50]	[ECCV'24]		55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9
VRS-HQ-7B [8]	[CVPR'25]		59.8	64.5	62.1	53.5	58.7	56.1	56.6	61.6	59.1
GLUS [20]	[CVPR'25]		58.3	56.0	60.7	48.8	53.9	51.4	-	-	-
HyperSeg [41]	[CVPR'25]		56.0	60.9	58.5	50.2	55.8	53.0	53.1	58.4	55.7
RGA3-3B [35]	[ICCV'25]		57.6	61.0	59.3	50.6	55.0	52.8	54.1	58.0	56.1
RGA3-7B [35]	[ICCV'25]		58.7	62.3	60.5	53.1	57.7	55.4	55.9	60.0	58.0
ViLLa [53]	[ICCV'25]		-	-	-	-	-	-	54.9	59.1	57.0
InstructSeg [42]	[ICCV'25]		54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5
CoT-RVS-online-7B [13]	[arXiv'25]	VLM-Based w/ Explicit Reasoning	-	-	-	-	-	-	43.5	48.8	46.2
CoT-RVS-offline-12B [13]	[arXiv'25]		-	-	-	-	-	-	43.4	50.9	47.1
ReVSeg-7B (Ours)	-		**63.3**	**68.1**	**65.7**	**55.4**	**61.8**	**58.6**	**59.3**	**65.0**	**62.1**

+11.2 points. The significant performance improvement demonstrates the effectiveness of ReVSeg with the proposed framework. Furthermore, given the zero-shot nature of ReasonVOS, these gains highlight our strong generalization and robustness under challenging open-world queries.

On ReVOS, we conduct a comprehensive comparison across nine metrics. Our ReVSeg-7B consistently ranks first, surpassing previous SOTAs, including several larger parameter systems, by a obvious margin. The across-the-board improvements on these reasoning VOS benchmarks substantiate the effectiveness of our explicit reasoning chain and the efficiency of the proposed training recipe.

***Referring* Video Object Segmentation.** As previous practice [2, 5, 14, 29, 35, 50, 53], we report the experiment results on three Ref-VOS benchmarks, i.e., Ref-YouTube-VOS, Ref-DAVIS17, and MeViS. As in Tab. 4, consistently, our ReVSeg-7B sets a new state of the art, improving $\mathcal{J}\&\mathcal{F}$ by +2.7 points on Ref-YouTube-VOS, +4.8 points on Ref-DAVIS17, and +8.5 points on MeViS against the previous SOTAs. Notably, MeViS is a motion-guided benchmark and is regarded as the most challenging Ref-VOS benchmark in GLUS [20]. Our substantial gains on MeViS indicate strong adaptability on complex video scenarios with intricate motion patterns. Although referring queries require less semantic reasoning than reasoning queries, our performance gains are still obvious and consistent. These results indicate ReVSeg offers stronger cross-modal video understanding, better temporal aggregation, and more accurate target object detection than prior art.

Zero-shot Reasoning *Image* Segmentation. Besides the main results, we ask whether post-training on *video* tasks improves spatial grounding that transfers to *image* reasoning segmentation. In other words, does our model truly learn better spatial grounding capabilities which is generalizable to other tasks? To probe this, we conduct a zero-shot evaluation on the reasoning image segmentation task. Specifically, we evaluate both the base model (Qwen2.5-VL-7B [1]) and our post-training ReVSeg-7B on ReasonSeg dataset [16], and report the mean per-image IoU (gIoU) and the cumulative IoU (cIoU) as in LISA [16]. As shown in Tab. 3, despite no image-specific training, ReVSeg yields delivers consistent gains on both test and validation splits compared to the base model. These improvements indicate that our pipeline no only simply couples spatial grounding with video understanding and temporal cues, but also upgrades its intrinsic spatial grounding ability.

Table 3. Zero-shot reasoning image segmentation results on ReasonSeg [16] dataset.

Method	test		val	
	gIoU	cIoU	gIoU	cIoU
Qwen2.5VL-7B	55.9	44.3	59.5	54.0
ReVSeg-7B (Ours)	**59.7**	**47.4**	**63.7**	**59.9**

Qualitative Results. Fig. 3 presents detailed case studies highlighting the role of explicit reasoning chains in VOS. In the first example, the video depicts a typical road scene without a visually salient target. The query asks what could triggered the driver's honk. The model parses the scene, integrates visual cues with traffic commonsense, and infers plausible causes, ultimately identifying the jaywalking pedestrian as the most likely trigger. Notably, the target object occupies only a small portion of the frame, making

Table 4. Video referring segmentation results on Ref-Youtube-VOS [29], Ref-DAVIS17 [14], MeViS [5] datasets.

Method		Type	Ref-YouTube-VOS			Ref-DAVIS17			MeViS		
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS [29]	[ECCV'20]	Segmentation Specialists	45.3	49.2	47.2	47.3	56.0	51.6	25.7	29.9	27.8
LBDT [6]	[CVPR'22]		48.2	50.6	49.4	-	-	54.1	27.8	30.8	29.3
MTTR [3]	[CVPR'22]		54.0	56.6	55.3	-	-	-	28.8	31.2	30.0
LMPM [5]	[ICCV'23]		-	-	-	-	-	-	34.2	40.2	37.2
ReferFormer [46]	[CVPR'22]		61.3	64.6	62.9	58.1	64.1	61.1	29.8	32.2	31.0
OnlineRefer [44]	[CVPR'23]		61.6	65.5	63.5	61.6	67.7	64.8	-	-	-
DsHmp [10]	[CVPR'24]		65.0	69.1	67.1	61.7	68.1	64.9	43.0	49.8	46.4
LISA-7B [16]	[CVPR'24]	VLM-Based w/ Latent Tokens	53.4	54.3	53.9	62.2	67.3	64.8	35.1	39.4	37.2
LISA-13B [16]	[CVPR'24]		54.0	54.8	54.5	63.2	68.8	66.9	35.8	40.0	37.9
TrackGPT-7B [55]	[arXiv'24]		55.3	57.4	56.4	59.4	67.0	63.2	37.6	42.6	40.1
TrackGPT-13B [55]	[arXiv'24]		58.1	60.8	59.5	62.7	70.4	66.5	39.2	43.1	41.2
VISA-7B [50]	[ECCV'24]		59.8	63.2	61.5	66.3	72.5	69.4	40.7	46.3	43.5
VISA-13B [50]	[ECCV'24]		61.4	64.7	63.0	67.0	73.8	70.4	41.8	47.1	44.5
VideoLISA [2]	[NeurIPS'24]		61.7	65.7	63.7	64.9	72.7	68.8	41.3	47.6	44.4
VRS-HQ-7B [8]	[CVPR'25]		68.3	72.5	70.4	72.6	79.4	76.0	47.6	53.7	50.6
GLUS [20]	[CVPR'25]		65.5	69.0	67.3	-	-	-	48.5	54.2	51.3
HyperSeg [41]	[CVPR'25]		-	-	68.5	-	-	71.2	-	-	-
RGA3-3B [35]	[ICCV'25]		65.8	69.1	67.4	67.6	76.6	72.1	46.2	51.5	48.8
RGA3-7B [35]	[ICCV'25]		66.8	70.1	68.5	68.3	77.3	72.8	47.4	52.8	50.1
ViLLa [53]	[ICCV'25]		64.6	70.4	67.5	70.6	78.0	74.3	46.5	52.3	49.4
InstructSeg [42]	[ICCV'25]		65.4	69.5	67.5	67.3	74.9	71.1	-	-	-
CoT-RVS-online-7B [13]	[arXiv'25]	VLM-Based w/ Explicit Reasoning	-	-	-	70.4	77.5	73.9	42.7	49.1	45.9
CoT-RVS-offline-13B [13]	[arXiv'25]		-	-	-	70.9	78.3	74.6	40.3	48.1	44.2
ReVSeg-7B (Ours)	-		**71.1**	**75.2**	**73.1**	**77.4**	**84.1**	**80.8**	**56.1**	**63.4**	**59.8**

Table 5. Ablation experiments for the proposed decoupling framework and RL post-training. The referring VOS results were evaluated on the Ref-DAVIS17 dataset, while the reasoning VOS results were evaluated on the ReasonVOS dataset. Decom. Pipeline indicates whether to use the decomposed pipeline, RL denotes the usage of Reinforcement Learning.

Decom. Pipeline	RL	Ref-DAVIS17			ReasonVOS		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
		3.2	5.6	4.4	7.4	8.8	8.1
	✓	4.0	6.0	5.3	7.1	8.1	7.6
✓		59.2	66.9	63.0	38.7	43.4	41.0
✓	✓	77.4	84.1	80.7	61.8	67.7	64.8

this case particularly challenging. In the second example, the model is queried about the creature posing the greatest threat. It first recognizes the most menacing species (elephant) and, guided by world knowledge that elder elephants protect calves and the herd, pinpoints the specific individual with high precision. Across cases, the reasoning chain consistently selects well-segmentable keyframes, stabilizing grounding and mitigating error accumulation, which yields cleaner masks and more consistent trajectories. Additional visualizations are provided in the supplementary.

4.3. Ablation Studies

Key Pipeline Design. We isolate the contributions of decomposed reasoning chain and RL post-training via four variants. (1) *Base Model*, which directly predicts spatio-temporal grounding from the video and query. (2) *Base Model + RL*, which applies GRPO post-training to the base model. (3) *Decomposed*, which restructures inference into our multi-turn decomposed reasoning chain. (4) *Decomposed + RL*, our final model ReVSeg with GRPO-enhanced reasoning. We evaluate them on Ref-DAVIS17 (in-domain Ref-VOS) and ReasonVOS (out-of-domain reasoning VOS), the results are showed in Tab. 5. The base model exhibits very poor video grounding ability, and RL alone fails to lift performance due to sparse effective rollouts and credit assignment in end-to-end prediction. In contrast, decomposed pipeline yields a clear jump by guiding the model to compose its primitive skills (video understanding, temporal grounding, spatial grounding) into a coherent procedure. Adding RL further drives self-evolution, tightening the interplay of these skills for VOS reasoning and delivering substantial additional gains across both datasets. Together, decomposed reasoning and RL post-training are necessary and complementary, culminating in unprecedented performance.

Frame Sampling Strategy. We examine the effect of input frame count during training and inference. As shown in Tab. 6, we train models with 12 / 16 / 20 uniformly sampled frames and observe the diminishing returns beyond 16. Accuracy improves from 12 to 16, while additional frames bring only marginal gains but increase training cost roughly. Balancing efficiency and performance, we adopt 16 frames as the default for both training and evaluation. At test time, performance remains stable across different frame counts, indicating that the reasoning pipeline is robust to temporal

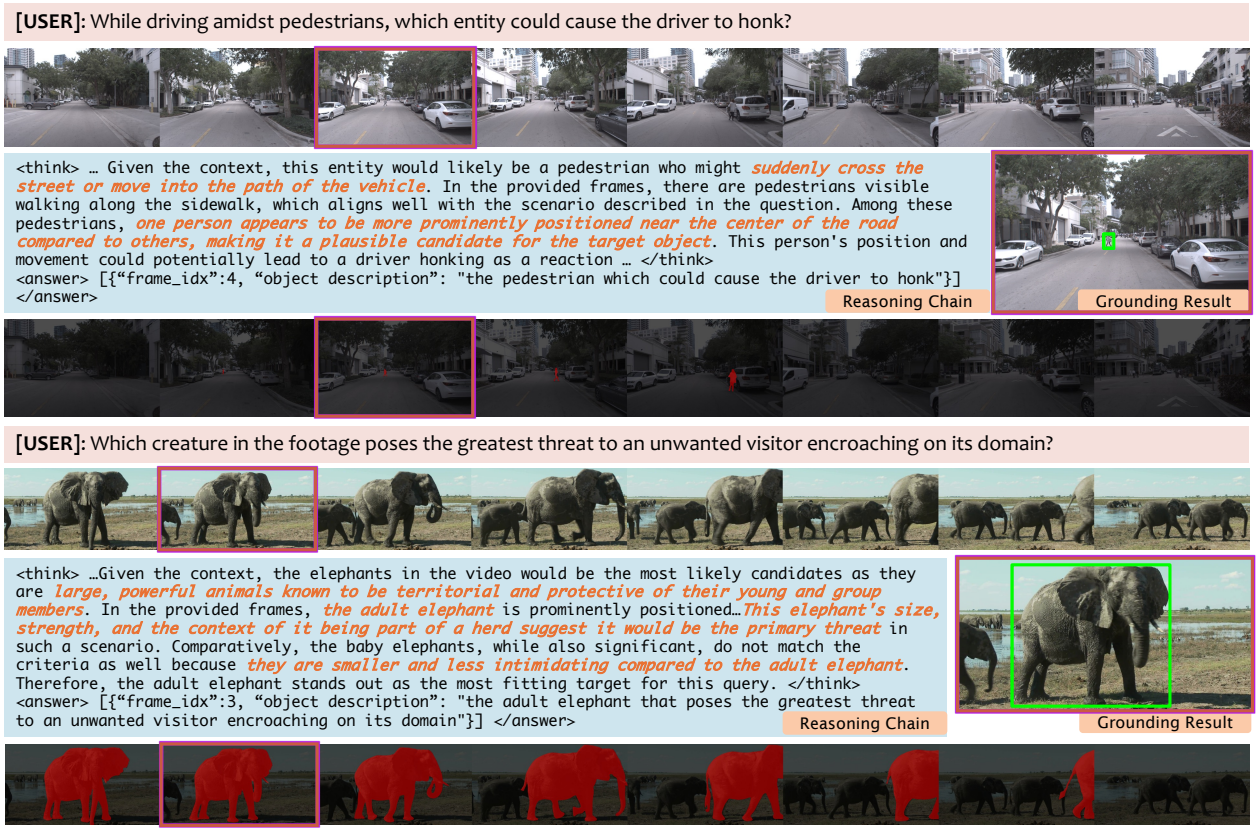

Figure 3. Qualitative cases of ReVSeg on ReasonVOS [2]. The frame highlighted in red indicates the selected keyframe. The green bounding box within the enlarged keyframe on the right size represents the grounding result. Zoom in to view visual details.

Table 6. Results obtained by employing different frame sampling rates during both the training and testing phases separately.

#Training Frames	#Testing Frames	Ref-DAVIS17 $\mathcal{J}$	Ref-DAVIS17 $\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	Training Time per Step (s)
12	12	76.8	83.8	80.3	603
16	16	**77.4**	**84.1**	**80.7**	725
20	20	77.0	**84.1**	80.5	831
16	12	76.6	83.9	80.3	
16	16	77.4	**84.1**	80.7	725
16	20	**77.5**	**84.1**	**80.8**	

Table 7. Model performance on VOS with three different temporal reward types. The referring VOS results were evaluated on the MeViS dataset, while the reasoning VOS results were evaluated on the ReasonVOS dataset.

Temporal Reward Type	MeViS $\mathcal{J}$	MeViS $\mathcal{F}$	MeViS $\mathcal{J}\&\mathcal{F}$	ReasonVOS $\mathcal{J}$	ReasonVOS $\mathcal{F}$	ReasonVOS $\mathcal{J}\&\mathcal{F}$
No Reward	50.7	58.7	54.7	55.0	61.0	58.0
0/1 Reward	53.6	61.2	57.4	56.7	62.7	59.7
Soft Reward	**56.1**	**63.4**	**59.8**	**61.8**	**67.7**	**64.8**

sampling density.

Designs of Reward. Temporal rewards are crucial for teaching the model to select keyframes that genuinely facilitate downstream spatial grounding and mask decoding. Prior keyframe selection works [12, 19, 49, 52] primarily score semantic relevance, which is insufficient for VOS: a good frame should also reveal the target with minimal occlusion and adequate scale. We therefore ablate three temporal reward schemes in RL: (1) no temporal reward, (2) a binary 0/1 reward that only checks whether the target appears in the selected frame(s), and (3) our soft temporal reward that scores object visibility via normalized bounding box area. As shown in Tab. 7, the soft variant provides a graded learning signal aligned with object visibility and scale, alleviates sparse credit assignment, and discourages degenerate selections (e.g., heavily occluded or tiny targets).

5. Conclusion

In this work, we reformulate reasoning-centric VOS as a sequential decision problem and introduce ReVSeg, which decomposes the task into three pretrained primitive capabilities. We further develop a reinforcement learning framework that directly optimizes the reasoning chain, enabling the model to refine its decision quality without relying on dense supervision. Experiments demonstrate state-of-the-art performance across multiple VOS bench-

marks and uncover interpretable, stepwise reasoning trajectories. We believe that this explicit-chain, outcome-driven formulation provides a general paradigm for advancing reasoning-aligned video understanding models.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5, 6

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 1, 2, 5, 6, 7, 8

[3] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 2, 5, 6, 7

[4] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 2

[5] 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, 5, 6, 7

[6] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4964–4973, 2022. 5, 7

[7] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 2

[8] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29183–29192, 2025. 1, 2, 5, 6, 7

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 4

[10] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13332–13341, 2024. 5, 7

[11] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025. 2, 3

[12] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712, 2025. 8

[13] Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Cot-rvs: Zero-shot chain-of-thought reasoning segmentation for videos. *arXiv preprint arXiv:2505.18561*, 2025. 2, 5, 6, 7

[14] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian conference on computer vision*, pages 123–141. Springer, 2018. 5, 6, 7

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[16] 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. 1, 2, 5, 6, 7

[17] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 2

[18] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 5, 6

[19] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*, 2024. 8

[20] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8658–8667, 2025. 1, 2, 5, 6, 7

[21] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 2, 4

[22] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 2, 4

[23] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc:

Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36:26425–26437, 2023. 5

[24] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 2

[25] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 5

[26] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 5

[27] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025. 2

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[29] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223. Springer, 2020. 1, 2, 5, 6, 7

[30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 4, 5

[31] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 3

[32] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 2, 3

[33] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. 4

[34] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, XU Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 5

[35] Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22274–22284, 2025. 1, 2, 5, 6, 7

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[37] Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, et al. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning. *arXiv preprint arXiv:2504.16656*, 2025. 2

[38] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3

[39] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 4

[40] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 1

[41] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, and Yujiu Yang. Hyperseg: Hybrid segmentation assistant with fine-grained visual perceiver. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8931–8941, 2025. 1, 2, 5, 6, 7

[42] Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20193–20203, 2025. 1, 2, 5, 6, 7

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 4

[44] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023. 5, 7

[45] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2(3): 4, 2021. 1, 2

[46] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 5, 6, 7

[47] Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025. 2

[48] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025. 2

[49] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 8

[50] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024. 1, 2, 5, 6, 7

[51] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv. org/abs/2305.10601*, 3:1, 2023. 4

[52] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. *arXiv preprint arXiv:2506.22139*, 2025. 8

[53] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. Villa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23667–23677, 2025. 1, 2, 5, 6, 7

[54] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 2

[55] 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. 2, 6, 7

Appendix

A. Training Curves

In Fig. 4, we summarize the dynamics of ReVSeg training. The first row plots the three reward components. Format reward r_f in (a) rises sharply to near-perfect score within the first few steps and remains saturated thereafter, indicating that the policy quickly masters the instructed output format and consistently completes the two-round rollout generations. Temporal reward r_t and spatial reward r_s in (b) and (c) follow a similar growth pattern early on, suggesting that initial gains are driven primarily by producing well-formed, parseable responses that expose the temporal-spatial grounding results. As the format reward plateaus, the growth of temporal and spatial rewards decelerates improvements, steming from stronger reasoning rather than formatting, reflecting better temporal selection and tighter localization.

The second row reports response length, total reward r and rollout turns. Response length in (d) increases during the early phase as the proportion of complete two round rollouts rises, and shows pronounced oscillations between roughly 100 and 250 steps. The analysis of training output reveals several shifts in response policy before settling into to a stable reasoning pattern. Total reward in (e) exhibits a steady increase over training. Rollout turns in (f), defined as the average number of rounds per sample within a batch, quickly converge to two, consistent with the observed trends of the reward components.

Overall, these curves indicate a well-behaved optimization, ReVSeg promptly adopts the intend two-round policy and continues to improve its reasoning quality in a stable manner.

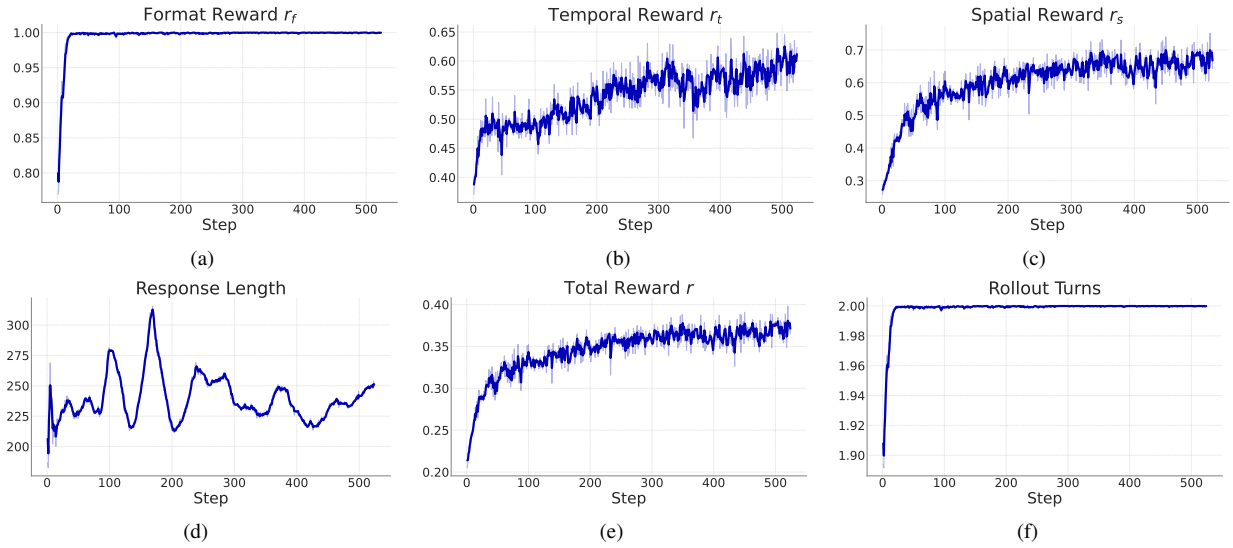

Figure 4. Training curves of ReVSeg. (a) Format reward r_f rapidly converges to a full score and remains saturated. (b) Temporal reward r_t and (c) Spatial reward r_s increase steadily with training. (d) Response length remains stable overall without collapse. (e) Total reward r rises consistently over time. (f) Average number of rollout turns quickly converge to 2.

B. Qualitative Results

In Fig. 5, we present more visualization examples highlighting the reasoning-centric segmentation capability of ReVSeg. Across diverse scenes, the model analyzes temporal dynamics, including object interactions (*e.g.*, the kittens interact with orange ballon, person holding green towel) and motion cues (*e.g.*, car moving in the opposite direction compared to the motorcycles), while aligning them with the abstract semantics of the user query. It integrates the video evidence with commonsense knowledge (*e.g.*, the concrete mixer truck designed for construction sites) to identify the target entity, selects frames that are favorable for downstream localization (*e.g.*, bright color, centrally located in the frames, visually against the background, size), and converts this selection into a keyframe index and a concrete object description. Conditioned on this description, ReVSeg produces tight spatial grounding and propagates masks across the video sequence, yielding the fine-grained segmentation masks.

Figure 5. Additional qualitative cases of ReVSeg. The frame highlighted in red indicates the selected keyframe. The green bounding box within the enlarged keyframe on the right side represents the grounding result. Zoom in to view visual details.

"Here is an {*second*}-second video represented by {*nframes*} frames (fps=2). Please first think deeply based on the given video, and then find '{*question*}' with temporal timestamp and detailed object description for spatial grounding.
Compare the difference between all possible objects and find the object that most matches the query.
Carefully identify all frames and select the keyframe that will minimize the effort for a segmentation model to find the target object. Prioritize frame where the target object is clearly visible and low occlusion, large enough in image area, visually distinct from background and other objects, minimally motion-blurred, and centered.
Describe the static appearance of the target object in the selected keyframe in detail, to aid subsequent spatial grounding.
The reasoning process and answer must be enclosed within <think> </think> and <answer> </answer> tags, respectively.
Output the keyframe index (0-{*nframes-1*}) and specific description of the target object in that frame in JSON format.
i.e., <think> reasoning process here </think>
<answer> [{'frame_idx': 7, 'object description': 'the red triangular stone on the far left'}] </answer>."

(a) User prompt for first round generation, where $\{second\}$, $\{nframes\}$ and $\{question\}$ refer to the video duration T/fps, the frame number T of input video sequence V and input query $\mathbf{x}$, respectively.

"This is the extracted keyframe, please detect '{*object description*}' with bbox coordinates in JSON format."

(b) User prompt for second round generation, where $\{object\ description\}$ refers to the extracted object description $\mathbf{d}$ from the first round response $\mathbf{y_1}$.

Figure 6. User prompt templates for two-round rollout.

C. Implementation Details

In this section, we supplement some implementation details, including user prompts in Sec. C.1 and pseudocode for our proposed decomposed generation with reasoning chain in Sec. C.2.

C.1. ReVSeg User Prompt Details

To elicit reliable reasoning for video object segmentation, we design a task-specific prompting scheme tailored to our two-round rollout of the policy VLM $\mathcal{F}$, as shown in Fig. 6. Each round uses a distinct prompt that mirrors the decomposition of the task.

Round one: user prompt for video understanding and temporal grounding. The first prompt follows a chain-of-thought style to encourage deep analysis. It asks the model to compare salient objects across the video, examine their occurrences over frames, and select a keyframe where the target is most suitable for localization. The prompt instructs $\mathcal{F}$ to first output a reasoning trace, then a structured result containing the keyframe index and a concise object description.

Round two: user prompt for spatial grounding. The second prompt provides the keyframe and object description extracted from round one response and requests precise localization on the selected frame. The output must follow a JSON format containing a tight bounding box.

Both prompts request reasoning separated from the final answer, standardized JSON fields to stabilize generation and reduce parsing errors. In practice, this design yields interpretable responses, that serve as robust visual prompts for subsequent mask propagation.

C.2. Algorithm

We distill the key generation component of ReVSeg into a compact pseudocode summarized in Algorithm 1. For each input video-query input, the model performs n times two-round rollout sampling and ultimately returns the resulting trajectories set $\{o_i\}_{i=1}^{n}$.

Algorithm 1: Decomposed Generation of ReVSeg with Two-Round Rollout

Require: Video $V = \{I_t\}_{t=1}^{T}$, query $\mathbf{x}$, policy VLM $\mathcal{F}$, parser $\mathcal{G}$, group size n, prompt template $\mathcal{P}_1$ and $\mathcal{P}_2$
Ensure : Rollout sequences $\{o_i\}_{i=1}^{n}$

for $i = 1$ **to** n **do**
  Initialize current VLM input sequence $\mathbf{z} \leftarrow \mathcal{P}_1(V, \mathbf{x})$
  Initialize current VLM rollout sequence $o_i \leftarrow \varnothing$
  `// Round-1: video understanding + temporal grounding`
  Generate the first round response sequence $\mathbf{y_1} \sim \mathcal{F}(\cdot \mid \mathbf{z})$
  Extract status flag $S_1 \in \{\texttt{succ}, \texttt{fail}\}$, keyframe index k and object description $\mathbf{d}$ through parser $(S_1, k, \mathbf{d}) \leftarrow \mathcal{G}(\mathbf{y_1})$
  Append $\mathbf{y}_1$ to input sequence $\mathbf{z} \leftarrow \mathbf{z} \oplus \mathbf{y}_1$
  Append $\mathbf{y}_1$ to rollout sequence $o_i \leftarrow o_i \oplus \mathbf{y}_1$
  if $S_1 = \texttt{fail}$ **then**
    Terminate rollout generation early and flag status $S_2 \leftarrow \texttt{fail}$, set keyframe I_k and bounding box B_k to Null $I_k, B_k \leftarrow \texttt{null}$
    Continue
  end
  `// Round-2: spatial grounding`
  Select the keyframe I_k from video $I_k \leftarrow V[k]$
  Append $(I_k, \mathbf{d})$ to the ongoing input sequence $\mathbf{z} \leftarrow \mathbf{z} \oplus \mathcal{P}_2(I_k, \mathbf{d})$
  Generate the second response sequence $\mathbf{y_2} \sim \mathcal{F}(\cdot \mid \mathbf{z})$
  Extract status flag $S_2 \in \{\texttt{succ}, \texttt{fail}\}$ and spatial grounding result $B_k \in \mathbb{R}^4_{\geq 0}$ through parser $(S_2, B_k) \leftarrow \mathcal{G}(\mathbf{y_2})$
  Append $\mathbf{y_2}$ to rollout sequence $o_i \leftarrow o_i \oplus \mathbf{y_2}$
end
return *final generated rollouts* $\{o_i\}_{i=1}^{n}$