

VideoSEG-O3: A Multi-turn Reinforcement Learning Framework for Reasoning Video Object Segmentation

Anonymous Authors¹

Abstract

Reasoning Video Object Segmentation (RVOS) demands a sophisticated integration of temporal dynamics, spatial details, and linguistic reasoning to achieve precise pixel-level localization. Existing methods are limited to reasoning over fixed initial inputs and lack the capacity to actively acquire further visual evidence, which is often essential for resolving complex references in long or intricate videos. To address this, we propose **VideoSEG-O3**, the first multi-turn reinforcement learning framework for RVOS that emulates the human “*coarse-to-fine*” cognitive process. It employs a *multi-turn temporal-spatial chain-of-thought* to capture fine-grained details by iteratively pinpointing critical intervals and keyframes. Additionally, to enable the policy to perceive segmentation quality beyond mere text probability of [SEG] during the RL stage, we introduce *SEG-aware logit calibration*, which integrates pixel-wise segmentation feedback directly into the token-level logits. Furthermore, we design a *decoupled thinking trace* to hierarchically decompose the reasoning process into temporal, spatial, and linguistic dimensions, and construct **VTS-CoT**, a specialized cold-start dataset featuring comprehensive reasoning trajectories. Extensive experiments demonstrate that VideoSEG-O3 achieves advanced performance across 8 mainstream RVOS benchmarks, particularly excelling in long-horizon and complex reasoning tasks.

1. Introduction

Reasoning Video Object Segmentation (RVOS) aims to localize and segment dynamic objects across video sequences

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

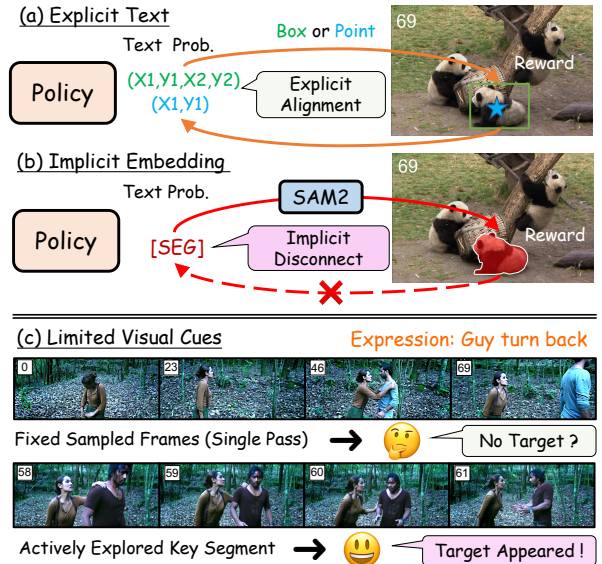


Figure 1. Motivation of **VideoSEG-O3**. (a) **Explicit Text**: Coordinate-based RL methods (e.g., box, point) allow rewards to directly optimize the policy. (b) **Implicit Embedding**: Latent mask representations (e.g., [SEG]) suffer from implicit disconnection between token probability and mask quality. (c) **Limited Visual Cues**: Fixed sampling strategies fail to actively explore key temporal segments required for precise localization.

based on complex natural language descriptions (Yan et al., 2024; Bai et al., 2024; Deng et al., 2025). Recently, Multi-modal Large Language Models (MLLMs) (Bai et al., 2025; Li et al., 2024a) have catalyzed progress in this field by leveraging their robust visual-language understanding capabilities. Existing MLLM-based solutions generally follow two trajectories: *implicit approaches* (Lai et al., 2024; Yuan et al., 2025a;b), which utilize latent embeddings (e.g., [SEG]) to guide mask decoders such as SAM (Ravi et al., 2024), and *explicit approaches* (Liu et al., 2025b; Huang et al., 2025; Lu et al., 2025; Zhu et al., 2025; Zhong et al., 2025; Gong et al., 2025a), which generate geometric prompts like bounding boxes or points to trigger external segmenters (Ravi et al., 2024; Cheng & Schwing, 2022).

Within the scope of RL applications, the majority of works (Zhu et al., 2025; Gong et al., 2025a) favor the explicit paradigm (Fig. 1(a)). This preference stems from two

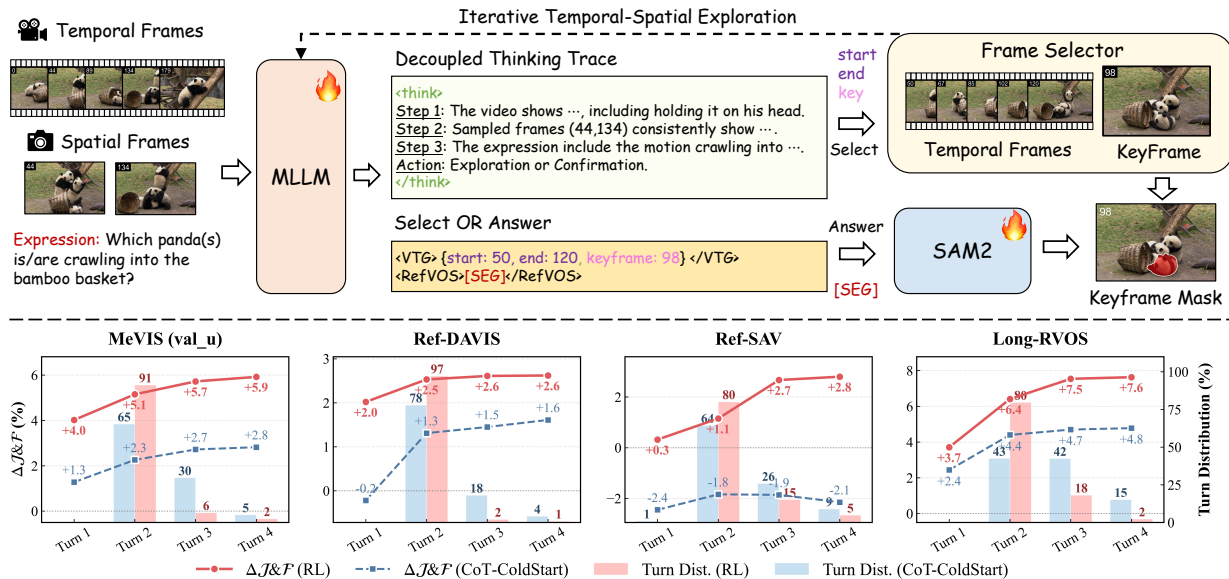


Figure 2. Overview and statistical analysis of **VideoSEG-O3**. **Top**: The MLLM architecture employs a *Decoupled Thinking Trace* to analyze motion context, spatial details, and expressions. A specialized `<select>` token is utilized to orchestrate iterative temporal-spatial exploration. **Bottom**: Performance gains ($\Delta\mathcal{J}\&\mathcal{F}$) and turn distributions across four representative benchmarks. Results highlight the performance evolution and adaptive reasoning behavior for both CoT cold-start and RL stages.

primary factors: *First*, MLLMs (Bai et al., 2025) possess an inherent proficiency in spatial grounding. *Second*, since geometric prompts are generated as text tokens, the policy can be optimized directly via token-level probabilities. However, the explicit paradigm faces some limitations. *First*, it is inherently indirect, as sparse geometric prompts lack the capacity to accurately characterize fine-grained instance. *Second*, boxes and points are frame-specific, making them insufficient to serve as stable, unified instance representations across a video. In contrast, implicit embeddings can provide consistent descriptors that maintain robust object identity throughout the sequence (Bai et al., 2024).

Furthermore, existing RVOS methods (Liu et al., 2025a; Feng et al., 2025) are restricted to reasoning over a fixed set of sampled frames, lacking the agency to actively seek supplementary visual evidence when the initial context is insufficient to resolve intricate references. As illustrated in Fig. 1(c), such fixed sampling strategies (Yuan et al., 2025a) inevitably overlook critical context, leading to localization failures. Moreover, these methods fail to utilize visual tokens efficiently. Specifically, conditioned on a specific textual query, the vast majority of visual tokens in a video are often irrelevant. Consequently, processing high-resolution inputs incurs substantial computational redundancy due to the calculation of these uninformative frames tokens.

To address these challenges, we propose **VideoSEG-O3**, which is governed by three pivotal imperatives: (1) Transitioning from static single-pass inference to a *multi-turn temporal-spatial Chain-of-Thought*. (2) Bridging the modal-

ity gap between discrete text generation and dense mask prediction by enabling *direct token-level optimization* for implicit latent embeddings (e.g., [SEG]); and (3) *Decoupling thinking process* into structured dimensions of temporal dynamics, spatial details, and linguistic expressions to facilitate a more granular understanding.

(1) **Multi-turn Temporal-Spatial Chain-of-Thought**. Diverging from single-pass segmentation frameworks, VideoSEG-O3 adopts an active spatiotemporal exploration architecture. As shown in Fig. 2 (Top), VideoSEG-O3 utilize the `<select>` token to orchestrate iterative exploration, dynamically integrating temporal intervals and spatial keyframe. This explicitly decouples temporal action understanding from fine-grained spatial localization. Crucially, the model autonomously determines the optimal reasoning depth, continuously gathering supplementary visual cues to refine the segmentation. Quantitative analysis in Fig. 2 (Bottom) further validates this design: (1) *Progressive Gain*: Performance consistently scales with interaction turns across both cold-start and RL stages, demonstrating the effectiveness of additional reasoning processes and the incorporation of supplementary visual cues. (2) *Reasoning Efficiency*: Post-RL, the model achieves superior performance with reduced interaction turns. This indicates that RL optimizes the select policy to precisely target decisive temporal frames. (3) *Adaptive Termination*: The model dynamically adjusts its search depth relative to task complexity. It converges in fewer turns on simple benchmarks (e.g., Ref-DAVIS) while allocating additional visual cues for intricate scenarios (e.g., Ref-SAV, LongRVOS).

(2) SEG-aware Logit Calibration. VideoSEG-O3 adopts the implicit latent embedding paradigm. However, standard GRPO (DeepSeek-AI, 2025) optimizes the likelihood of the [SEG] token but overlooks the quality of its latent embedding required for accurate segmentation (Fig. 1(b)). To bridge this gap, we introduce *SEG-aware Logit Calibration*, which modulates the [SEG] token’s logits using pixel-wise segmentation probabilities. This ensures that the RL process synchronously optimizes both the textual reasoning path and the quality of the segmentation embedding.

(3) Decoupled Thinking Trace. Effective RVOS necessitates the synergistic comprehension of temporal dynamics, spatial details, and linguistic semantics. Unlike prior works that rely on holistic but entangled analysis (Zhong et al., 2025; Zhu et al., 2025), we design a *Decoupled Thinking Trace* that hierarchically decomposes the reasoning process into three structured stages, supported by resolution-adaptive visual inputs. As shown in Fig. 2 (Top), the MLLM ingests low-resolution frames to capture global temporal context, while utilizing high-resolution keyframes for spatial precision. The reasoning trajectory is strictly organized into: (i) *Temporal Understanding*, (ii) *Spatial Detail Capturing*, and (iii) *Expression Parsing*.

In summary, our key contributions are as follows: **(1)** We propose **VideoSEG-O3**, the first multi-turn RL framework for RVOS. By incorporating a *Multi-turn Temporal-Spatial CoT*, it shifts the paradigm from passive one-step inference to iterative visual exploration, enabling the model to autonomously perceive fine-grained details. **(2)** We introduce *SEG-Aware Logit Calibration*, which injects pixel-wise segmentation feedback directly into the [SEG] token’s policy. **(3)** We construct **VTS-CoT**, a cold-start dataset featuring structured reasoning trajectories across temporal, spatial, and linguistic dimensions. **(4)** **VideoSEG-O3** achieves advanced performance across 8 RVOS benchmarks. Remarkably, even with a compact model scale (4B), it surpasses previous SOTA methods, achieving gains of +6.1% on the long-video task LongRVOS, +4.2% on the motion-centric MeViS, and +4.0% on the reasoning benchmark ReVOS.

2. Related Works

Supervised Fine-Tuning Approaches. SFT serves as the dominant paradigm for adapting MLLMs to RVOS (Zhu et al., 2023; Zheng et al., 2024; Lin et al., 2025; Kao et al., 2025; Munasinghe et al., 2024). Pioneering efforts, such as VISA (Yan et al., 2024), combine MLLMs for temporal reasoning with external object trackers (Cheng & Schwing, 2022) for mask propagation. VideoLISA (Bai et al., 2024) advances this by introducing a sparse-dense sampling strategy and a *One-Token-Seg-All* paradigm. More recently, VRS-HQ (Gong et al., 2025b) enhances temporal consistency via token fusion and occlusion-aware keyframe selec-

tion using SAM2 (Ravi et al., 2024), while MomentSeg (Dai et al., 2025c) improves RVOS performance by incorporating temporal sentence grounding capabilities to refine keyframe identification. Furthermore, Sa2VA and UniPixel (Yuan et al., 2025a; Liu et al., 2025a) establish a unified framework for multimodal dense localization and visual-prompt-based captioning tasks, achieving significant performance improvements. However, these methods lack an explicit reasoning trace to bridge the linguistic-visual gap and often neglect joint spatiotemporal reasoning. To address this, we propose **VideoSEG-O3**, which systematically integrates a *Multi-turn Temporal-Spatial Chain-of-Thought* to enhance both reasoning depth and segmentation quality.

Reinforcement Learning Approaches. RL has established itself as a potent paradigm for enhancing reasoning capabilities in LLMs, particularly via GRPO (DeepSeek-AI, 2025), and has recently seen preliminary exploration in reasoning segmentation tasks. In the **image domain**, existing methods (Liu et al., 2025c; Lu et al., 2025; Huang et al., 2025; Wang et al., 2025) leverage specific rewards to improve both image-level reasoning and segmentation quality. Representative approaches like SegZero (Liu et al., 2025b) adopt a box-mediated paradigm, utilizing predicted bounding boxes to prompt the SAM model for segmentation. Conversely, LENS (Zhu et al., 2025) introduces learnable tokens to facilitate end-to-end optimization, while R-Sa2VA (Yuan et al., 2025b) retains the [SEG] token paradigm (Lai et al., 2024) to trigger segmentation via a latent embedding. Transitioning to the **video domain**, recent works (Zhong et al., 2025; Gong et al., 2025a) instruct LLMs to generate explicit textual prompts (e.g., bounding boxes or points) via a Chain-of-Thought process, which subsequently guide the SAM2 for mask propagation. However, these methods overlook the potential of multi-turn interaction and supplementary visual cues for iterative refinement. To address this, **VideoSEG-O3** introduces autonomous *Multi-turn Spatiotemporal Exploration* and *SEG-aware Logit Calibration*, enabling direct optimization of the implicit embedding.

3. VideoSEG-O3

3.1. Task Formulation

In this work, we formulate VideoSEG-O3 as a sequential decision-making process that emulates human visual cognition, iteratively attending to fine-grained video temporal-spatial fragments to achieve precise segmentation. The optimization framework targets two core capabilities: (1) *Active Spatiotemporal Exploration*, enabling the model to identify informative temporal intervals and spatial keyframes; and (2) *Implicit Representation Refinement*, optimizing the [SEG] token embedding for high-fidelity mask generation. Specifically, the reasoning process begins with a global perception phase, where the model analyzes the full video

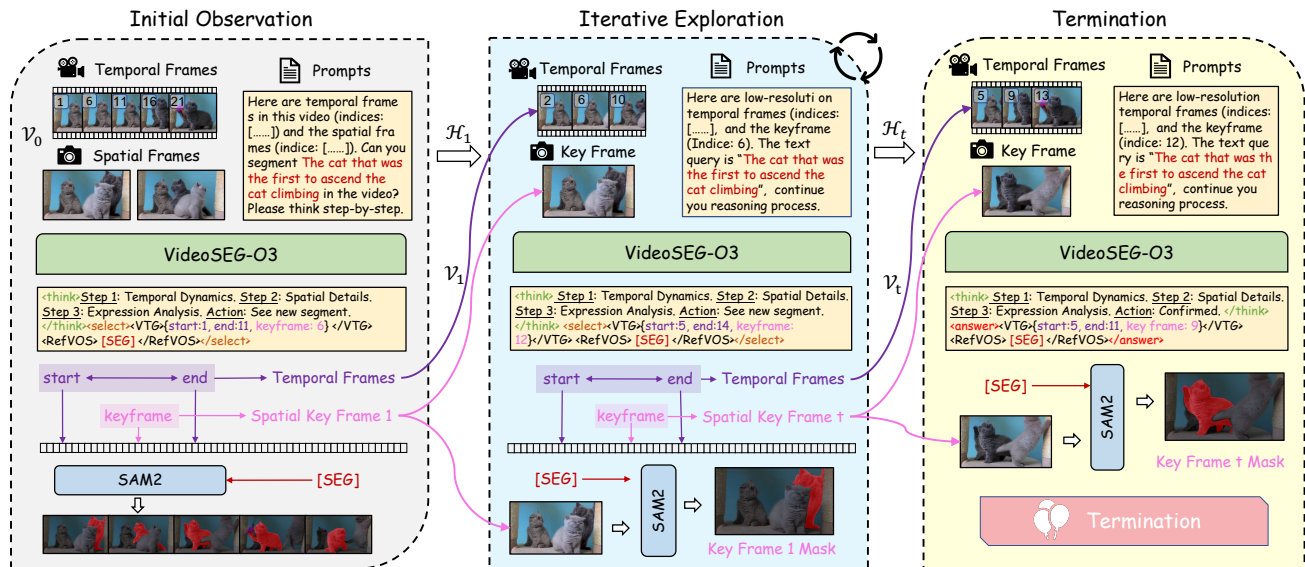


Figure 3. Overall pipeline of VideoSEG-O3. (1) **Initial Observation**: The model ingests global temporal frames and uniformly sampled spatial frames as visual inputs. (2) **Iterative Exploration**: The model iteratively refines its reasoning by taking the temporal interval and keyframe index selected via `<select>` in the previous turn as input. (3) **Termination**: Upon generating the `<answer>` token, the process terminates, and the `[SEG]` token yields the segmentation mask for the target keyframe.

context alongside sparse spatial cues. This global understanding guides the generation of a `<select>` token, triggering the active selection of critical temporal intervals and keyframes for detailed inspection. In subsequent iterations, the visual input is dynamically augmented with these refined spatiotemporal details. This cycle continues until the model determines that sufficient evidence has been gathered, signaling termination via the `<answer>` token to produce the final segmentation mask. As illustrated in Fig. 3, this process is formally defined as a Markov Decision Process (MDP) specified by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$:

- **State** $s \in \mathcal{S}$: Represents the evolving observational status. Unlike traditional approaches that process a static video representation, our state is dynamic. We define the state at step t as a tuple $s_t = (\mathcal{V}_t, Q, \mathcal{H}_{t-1})$, where Q is the textual query and \mathcal{H}_{t-1} denotes the interaction history context. Crucially, \mathcal{V}_t represents the current visual observation: at $t = 0$, \mathcal{V}_0 comprises globally sampled frames for initial context; for $t > 0$, \mathcal{V}_t is actively updated to include *low-resolution temporal frames* and a *high-resolution keyframe*, derived from the previous action.
- **Action** $a \in \mathcal{A}$: Refers to the decision made by the Policy. The action space consists of two distinct operations: (1) *Exploration*, where the model outputs a `<select>` token followed by a specific time range and keyframe index to request refined visual details; and (2) *Termination*, where the model outputs an `<answer>` token followed by the `[SEG]` token to generate the final segmentation mask.
- **Transition Function** $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$: Defines the update mechanism for the observation state. If a_t is an explo-

ration action, \mathcal{T} acts as a sampler that retrieves data from the raw video. Specifically, it extracts the requested temporal segment at *low resolution* to maintain motion context, while simultaneously extracting the target keyframe at *high resolution* to capture fine-grained spatial details, forming the new visual input \mathcal{V}_{t+1} for the next iteration.

- **Reward Function** $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: Evaluates the quality of the reasoning and segmentation. The reward function incorporates multiple components, including format, temporal, segmentation, and progressive reward (Sec. 3.3).

3.2. SEG-aware Logit Calibration

Standard GRPO (DeepSeek-AI, 2025) optimizes the policy π_θ via an importance sampling objective derived exclusively from textual token sequences. For a sequence $y = \{y_1, \dots, y_T\}$, the importance ratio at token t is typically defined solely on the text distribution:

$$\log r_t(\theta) = \log P_{\text{LLM}}^\theta(y_t) - \log P_{\text{LLM}}^{\theta_{\text{old}}}(y_t). \quad (1)$$

While effective for linguistic alignment, this formulation presents a structural limitation in segmentation tasks. Standard GRPO treats the `[SEG]` token identically to discrete textual tokens, deriving the policy gradient strictly from language head logits. This creates a *gradient disconnect*: the mask decoder is differentially detached from the policy’s estimation. Moreover, as illustrated in Fig. 1(b), the prediction probability of the `[SEG]` token is irrelevant to the reward derived from mask quality. To address this misalignment, we propose *SEG-aware Logit Calibration*, as illustrated in Fig. 4(c). We reformulate the segmentation process as a

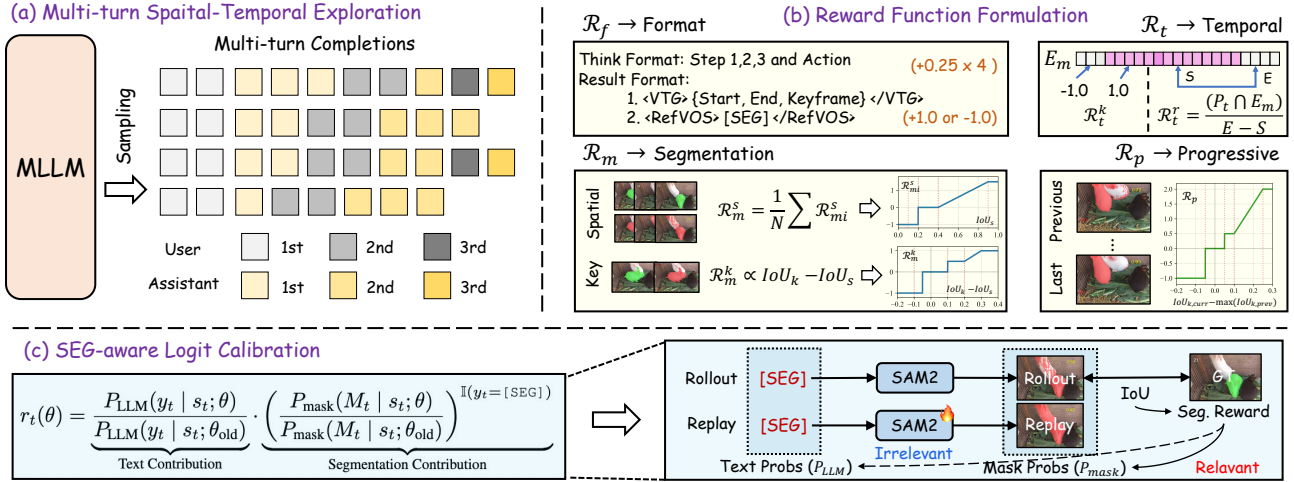


Figure 4. The overall framework of VideoSEG-O3. (a) Multi-turn temporal-spatial exploration (Sec. 3.1). (b) A composite reward design guiding reinforcement learning via format (\mathcal{R}_f), temporal (\mathcal{R}_t), segmentation (\mathcal{R}_m), and progressive (\mathcal{R}_p) rewards (Sec. 3.3). (c) A calibration strategy that aligns the latent [SEG] token representation with pixel-level mask confidence (Sec. 3.2).

composite probabilistic event that injects mask confidence into textual probabilities. Specifically, we calibrate the policy’s probability for the [SEG] token by directly injecting the pixel-wise confidence from the mask decoder:

$$\log \tilde{\pi}_{\theta}(y_t, M_t | s_t) = \log P_{LLM}(y_t | s_t) + \mathbb{I}(y_t = [SEG]) \cdot \log P_{mask}(M_t | \mathbf{h}_t). \quad (2)$$

To explicitly quantify the segmentation probability, we treat the binarized mask M_t obtained during the rollout phase as the sampled segmentation outcome. We define $\log P_{mask}$ as the spatial average of the log-likelihood that the mask decoder generates this specific rollout mask. Crucially, this probability is conditioned on \mathbf{h}_t , the hidden state representation of the [SEG] token at step t :

$$P_{mask}(M_t | \mathbf{h}_t) = \frac{1}{HW} \sum_{i,j} \mathcal{P}(m_{i,j} | \Phi(\mathbf{h}_t)_{i,j}). \quad (3)$$

Here, $m_{i,j} \in \{0, 1\}$ represents the pixel value from M_t , and $\Phi(\mathbf{h}_t)$ denotes the predicted mask probabilities projected from the hidden states \mathbf{h}_t . \mathcal{P} is computed via negative binary cross-entropy, effectively measuring the consistency between the model’s replay and rollout processes.

By substituting Eq. 2 into the GRPO objective, the importance ratio $r_t(\theta)$ becomes sensitive to visual fidelity. The log-ratio decomposes as:

$$\log r_t(\theta) \approx \underbrace{\left(\log P_{LLM}^{\theta}(y_t) - \log P_{LLM}^{\theta_{old}}(y_t) \right)}_{\text{Text Contribution}} + \underbrace{\mathbb{I}(y_t = [SEG]) \left(\log P_{mask}^{\theta}(M_t) - \log P_{mask}^{\theta_{old}}(M_t) \right)}_{\text{Segmentation Contribution}}. \quad (4)$$

This formulation establishes a pathway from the reward signal to the mask decoder. The gradient update can be mod-

ulated by the relative improvement in pixel-wise certainty. This ensures the policy is penalized not only for textual semantic drift but also for spatial uncertainty.

3.3. Reward and Objectives

Reward Functions Formulation. To incentivize structured reasoning and high-fidelity grounding, we formulate a composite reward, as illustrated in Fig. 4(b). Detailed descriptions are provided in Appendix C.1. We adopt an episodic reward, and the corresponding ablation studies and detailed analyses are presented in Appendix E.4.

- **Format Reward (\mathcal{R}_f):** We enforce strict compliance in two aspects: *Think Format* (ensuring logical CoT steps) and *Result Format* (validating syntactic correctness).
- **Temporal Reward (\mathcal{R}_t):** We evaluate localization accuracy through keyframe alignment and temporal precision:

$$\mathcal{R}_t^k \propto (-1)^{\mathbb{I}(k \notin G_t)}, \quad \mathcal{R}_t^r \propto \frac{|P_t \cap G_t|}{|P_t|}, \quad (5)$$

where k , P_t , and G_t denote the predicted keyframe, predicted range, and ground truth, respectively.

- **Segmentation Quality Reward (\mathcal{R}_m):** We employ reward that correlate with mask quality and keyframe selection superiority. Specifically, \mathcal{R}_m^s represents the cumulative frame-level segmentation quality, and \mathcal{R}_m^k quantifies the relative advantage of selected keyframe over manually sampled spatial frames, formulated as:

$$\mathcal{R}_m^s \propto \frac{1}{N} \sum_{i=1}^N IoU_i, \quad \mathcal{R}_m^k \propto (IoU_k - IoU_s). \quad (6)$$

- **Progressive Reward (\mathcal{R}_p):** To incentivize continuous refinement and penalize redundant interactions, we reward

the policy only when the current mask quality exceeds the historical maximum. Conversely, a penalty is applied for ineffective turns, formulated as:

$$\mathcal{R}_p \propto \left(\text{IoU}_{k,t} - \max_{j < t} (\text{IoU}_{k,j}) \right). \quad (7)$$

RL Training Objectives. To ensure stable policy optimization and prevent the degradation of segmentation capabilities during exploration, we formulate a hybrid objective function. Relying exclusively on sparse rewards can induce optimization instability, particularly when sampled generations yield flat reward distributions (i.e., vanishing advantage). To mitigate this, we introduce an auxiliary dense supervision signal. Analogous to the KL-divergence constraint in RLHF, this term strictly anchors the latent [SEG] representation to the GT masks. The total loss is defined as:

$$\mathcal{L}(\theta) = \mathcal{L}_{GRPO}(\tilde{\pi}_\theta) + \lambda_{seg} \underbrace{(\mathcal{L}_{spatial} + \mathcal{L}_{key})}_{\text{Joint Segmentation Loss}}, \quad (8)$$

where \mathcal{L}_{GRPO} maximizes the expected reward using the calibrated policy $\tilde{\pi}_\theta$. The auxiliary segmentation loss comprises $\mathcal{L}_{spatial}$, computed on the spatial frames, and \mathcal{L}_{key} , computed on the keyframe. This dual-objective mechanism ensures that the model maintains robust pixel-level decoding capabilities even during the early stages of reasoning exploration. The λ_{seg} balances the trade-off between semantic reasoning exploration and spatial grounding stability.

3.4. Multi-turn Temporal-Spatial Chain-of-Thought

VTS-CoT Dataset. We construct the cold-start CoT dataset through a four-stage pipeline: (1) *Data selection*, (2) *Temporal labeling*, (3) *Candidate generation*, and (4) *Decoupled CoT synthesis*, resulting in more than 6K samples. For the detailed construction process and the pipeline illustration, please refer to Appendix B.

Decoupled Thinking Trace. To facilitate structured reasoning, we design a *Decoupled Thinking Trace* that hierarchically organizes inference into three stages (Fig. 2, Top): (1) *Temporal Understanding* using global low-resolution frames, (2) *Spatial Description* via high-resolution spatial frames or keyframe, and (3) *Expression Parsing*. This decoupled design and resolution-adaptive workflow ensure the efficient utilization of input tokens.

4. Experiments

4.1. Implementation Details

Architecture and Setup. We construct our baseline by integrating Qwen3-VL (Bai et al., 2025) with the SAM2 (Ravi et al., 2024). Following prior approaches (Yuan et al.,

2025a), segmentation masks are generated by decoding the hidden state of the special [SEG] token via the SAM2 decoder. As outlined in Table 5, the training pipeline of VideoSEG-O3 comprises three distinct stages. During the SFT and CoT stages, we employ LoRA (Hu et al., 2021) to fine-tune the MLLM with a maximum sequence length of 8,192 tokens. Conversely, the RL stage involves full-parameter fine-tuning of the MLLM. All experiments are conducted on 8 NVIDIA H20 GPUs. The RL stage of VideoSEG-O3 involves 1,000 training steps. This process requires approximately 24 hours for VideoSEG-O3-4B and roughly 16 hours for the VideoSEG-O3-2B variant.

Training Stages. We adopt a hierarchical three-stage pipeline to progressively build VideoSEG-O3’s capabilities (detailed dataset composition is in Table 5). *Stage I (Supervised Fine-Tuning):* This phase establishes core competencies by aligning the [SEG] token with the mask decoder and, crucially, enforcing frame-level *temporal index* understanding to support discrete temporal reasoning. *Stage II (CoT Cold-Start):* Utilizing our constructed VTS-CoT dataset, this stage elicits structured multi-step reasoning and tool usage, standardizing the format for multi-turn interactions. *Stage III (Reinforcement Learning):* Finally, we optimize the policy for iterative exploration, leveraging composite rewards to guide the model in seeking discriminative visual cues and refining segmentation accuracy across turns.

4.2. Main Results

Referring Video Object Segmentation. Table 1 demonstrates that VideoSEG-O3-4B achieves state-of-the-art performance across five RefVOS benchmarks by a substantial margin. It exhibits superior generalization over specialized methods (Cuttano et al., 2025; Liang et al., 2025b) and dominates MLLMs-based methods (Yuan et al., 2025a) on complex tasks, achieving gains of **+15.5%** on Ref-SAV and **+6.1%** on Long-RVOS. On traditional benchmarks such as MeViS, Ref-YouTube-VOS, and Ref-DAVIS17, VideoSEG-O3 at a 2B scale achieves performance comparable to the previous SOTA UniPixel-7B (Liu et al., 2025a), demonstrating significantly higher parameter efficiency.

Reasoning Video Object Segmentation. As shown in Table 2, VideoSEG-O3 achieve advanced performance in both in-domain and out-domain scenarios. On ReVOS, it surpasses the strongest RL competitor, Veason-R1 (Gong et al., 2025a), by **+6.4%** overall, with a notable **+6.1%** gain in the *Reasoning* sub-category, validating the efficacy of our VideoSEG-O3 in handling complex linguistic logic. Furthermore, VideoSEG-O3-4B demonstrates robust zero-shot generalization, outperforming the MoRA-13B (Deng et al., 2025) on GroundMoRe benchmark by **+8.8%**.

Table 1. Main results on referring video object segmentation benchmarks. We compare **VideoSEG-O3** with specialized RVOS and MLLM-based methods. Best results are **bolded**, and the second best results are underlined.

Method	MeViS			Ref-Youtube-VOS			Ref-DAVIS17			Ref-SAV			Long-RVOS
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$
Specialized Methods													
SAMWISE (Cuttano et al., 2025)	49.5	46.6	52.4	69.2	67.8	70.6	70.6	67.4	74.5	–	–	–	35.6
ReferDINO (Liang et al., 2025b)	49.3	44.7	53.9	69.3	67.0	71.5	68.9	65.1	72.9	–	–	–	48.7
ReferMo (Liang et al., 2025a)	–	–	–	–	–	–	–	–	–	–	–	–	51.3
MLLM-based Methods													
LISA-7B (Lai et al., 2024)	37.2	35.1	39.4	53.9	53.4	54.3	64.8	62.2	67.3	–	–	–	–
VideoLISA-3.8B (Bai et al., 2024)	44.4	41.3	47.6	63.7	61.7	65.7	68.8	64.9	72.7	–	–	–	33.1
VISA-7B (Yan et al., 2024)	43.5	40.7	46.3	61.5	59.8	63.2	69.4	66.3	72.5	11.8	13.2	11.3	–
GLUS-7B (Lin et al., 2025)	51.3	48.5	54.2	67.3	65.5	69.0	–	–	–	–	–	–	36.6
InstructSeg-3B (Wei et al., 2025b)	–	–	–	67.5	65.4	69.5	71.1	67.3	74.9	–	–	–	–
Sa2VA-8B (Yuan et al., 2025a)	46.2	–	–	70.0	–	–	70.0	–	–	50.0	48.3	51.7	–
UniPixel-7B (Liu et al., 2025a)	<u>55.8</u>	<u>53.2</u>	58.3	<u>71.0</u>	<u>69.5</u>	72.4	76.4	72.7	80.1	–	–	–	–
VideoSEG-O3-2B (Ours)	55.6	52.5	<u>58.6</u>	70.5	68.3	<u>72.7</u>	80.0	76.0	84.0	<u>62.9</u>	<u>61.8</u>	<u>64.0</u>	<u>54.8</u>
VideoSEG-O3-4B (Ours)	60.0	57.0	63.0	74.1	71.5	76.6	<u>79.4</u>	<u>75.5</u>	<u>83.2</u>	65.5	64.3	66.7	57.4

Table 2. Main results on reasoning video object segmentation. We compare **VideoSEG-O3** with state-of-the-art SFT-based and RL-based methods. Our model establishes advanced performance across both in-domain (ReVOS) and out-domain (ReasonVOS, GroundMoRe) zero-shot benchmarks. Best results are **bolded**, and the second best results are underlined.

Method	ReVOS (In-Domain)									Out-Domain (ZS)	
	Referring			Reasoning			Overall			ReasonVOS	GroundMoRe
	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}		
SFT-based Methods											
VISA-13B (Yan et al., 2024)	57.4	55.6	59.1	44.3	42.0	46.7	50.9	48.8	52.9	–	5.31
VideoLISA-3.8B (Bai et al., 2024)	–	–	–	–	–	–	–	–	–	47.5	–
HyperSeg-3B (Wei et al., 2025a)	58.5	56.0	60.9	53.0	50.2	55.8	55.7	53.1	58.4	–	–
VRS-HQ-7B (Gong et al., 2025b)	62.1	59.8	64.5	56.1	53.5	58.7	59.1	56.6	61.6	–	–
MoRA-13B (Deng et al., 2025)	–	–	–	–	–	–	–	–	–	–	23.1
UniPixel-7B (Liu et al., 2025a)	65.8	63.9	67.8	61.5	<u>59.4</u>	63.7	63.7	61.7	65.7	–	–
RL-based Methods											
Veason-R1-7B (Gong et al., 2025a)	63.6	60.7	66.5	59.0	55.8	62.2	61.3	58.2	64.4	59.9	–
VideoSeg-R1-7B (Xu et al., 2025)	–	–	–	–	–	–	61.1	58.2	64.0	–	–
VideoSEG-O3-2B (Ours)	<u>67.5</u>	<u>65.0</u>	<u>70.0</u>	<u>62.0</u>	59.0	<u>65.0</u>	<u>64.8</u>	<u>62.0</u>	<u>67.5</u>	<u>60.2</u>	<u>29.1</u>
VideoSEG-O3-4B (Ours)	70.3	67.7	72.9	65.1	62.4	67.9	67.7	65.0	70.4	62.9	31.9

4.3. Ablation Study

Impact of Training Stages. Table 3 delineates the performance trajectory across our three-stage pipeline. **Stage I (SFT)** establishes the foundational grounding proficiency, achieving an average $\mathcal{J}\&\mathcal{F}$ of 54.96% by enabling the MLLM to generate [SEG] embeddings for the SAM2 decoder. Building upon this, **Stage II (CoT)** leverages structured reasoning and decoupled visual sampling to enhance spatiotemporal localization. This results in significant absolute gains on complex benchmarks, such as MeViS (val_u) (+2.72%) and Long-RVOS (+4.70%), raising the overall average to 57.07%. Although a performance fluctuation is observed on Ref-SAV (−1.86%) due to the distributional shift in VTS-CoT data, **Stage III (RL)** successfully recovers this drop with a substantial +6.53% improvement. Notably, the RL stage not only achieves an improved average $\mathcal{J}\&\mathcal{F}$

of 60.29% but also optimizes inference efficiency, reducing the *Avg. Round* by 20.1% (2.79 → 2.23). This demonstrates that the VideoSEG-O3 achieves superior performance with fewer interaction rounds, indicating a learned ability to precisely sample keyframes essential for target identification.

Refinement of Keyframe Selection via RL. Fig. 5 illustrates the comparative impact of reinforcement learning on the model’s keyframe selection capability across four benchmarks. We evaluate this refinement through two primary metrics: *keyframe mIoU*, which assesses the model’s ability to identify salient frames to provide high-quality initializations for SAM 2 propagation, and *keyframe empty ratio*, which measures the frequency of selecting frames where the target object is absent. Our results demonstrate that the RL-refined model achieves consistently higher mIoU scores across all datasets, signifying a superior capacity

Table 3. Ablation study on training stages. All metrics are reported using the $\mathcal{J}\&\mathcal{F}$ score. Incremental gains represent absolute improvements over the previous stage, except for *Avg. Round*, which denotes the percentage reduction. Evaluation is performed using keyframe initialization without post-processing.

Model	MeViS (val_u)	MeViS (val)	Ref-Youtube-VOS	ReVOS	Long-RVOS	Ref-SAV	Avg. $\mathcal{J}\&\mathcal{F}$	Avg. Rounds
VideoSEG-SFT	56.71	47.04	67.80	59.10	47.12	51.97	54.96	/
VideoSEG-CoT	59.43 (+2.72)	50.55 (+3.51)	70.12 (+2.32)	60.37 (+1.27)	51.82 (+4.70)	50.11 (-1.86)	57.07 (+2.11)	2.79
VideoSEG-RL	62.43 (+3.00)	53.22 (+2.67)	72.52 (+2.40)	64.24 (+3.87)	54.66 (+2.84)	54.64 (+4.53)	60.29 (+3.22)	2.23 (\downarrow 20.1%)

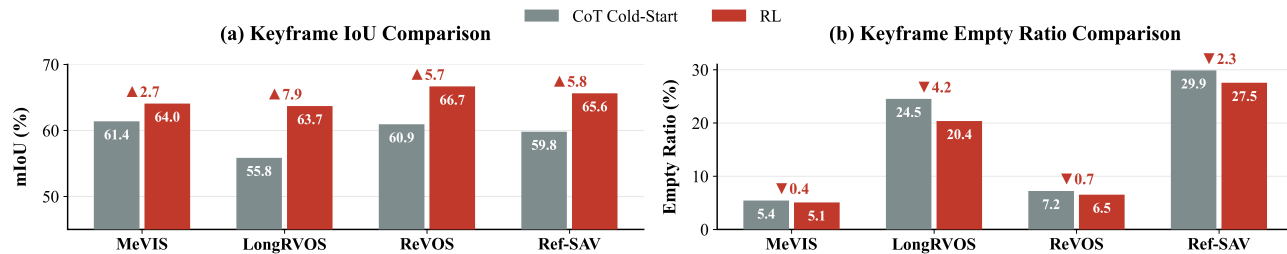


Figure 5. Quantitative comparison of RL on keyframe selection. We compare *keyframe mIoU* (mask quality for initialization) and *keyframe empty ratio* (frequency of selecting frames without the target) across four representative benchmarks.

Table 4. Ablation study on RL optimization components. All metrics are reported using the $\mathcal{J}\&\mathcal{F}$ score. All evaluations are conducted without post-processing to maintain consistency.

Model	MeViS (val_u)	Ref-SAV
Stage II (Cold-Start)	59.43	50.55
Baseline (GRPO-RL)	57.92 (-1.51)	49.78 (-0.77)
+ SEG-aware Logit Calibration	60.51 (+2.59)	51.27 (+1.49)
<i>Auxiliary Segmentation Loss</i>		
+ Spatial-wise	60.71 (+0.20)	51.76 (+0.49)
+ Keyframe-wise	61.31 (+0.80)	52.61 (+1.34)
+ Joint Segmentation Loss	61.95 (+1.44)	53.22 (+1.95)

for localizing discriminative visual anchors. Furthermore, a substantial reduction in the keyframe empty ratio post-RL indicates that the model autonomously learns to avoid non-informative frames, favoring target-salient views that effectively bridge the gap between linguistic expressions and dense pixel-level grounding.

Effectiveness of RL Optimization Components. We investigate the contribution of each RL component in Table 4. Our findings highlight three key insights: (1) Applying standard GRPO solely for text-level optimization leads to performance degradation (59.43% \rightarrow 57.92% on MeViS) compared to the Stage II baseline. Fundamentally, the gradients for the SAM2 decoder are obstructed because standard GRPO optimizes the probability of textual token generation. Consequently, an intrinsic disconnect arises between the segmentation-based reward and the text prediction, preventing the reward signal from being effectively attributed to the rollout samples. (2) The introduction of *SEG-aware Logit Calibration* successfully reverses this trend by aligning the

optimization objective with pixel-level fidelity, yielding immediate gains (49.78% \rightarrow 51.27% on Ref-SAV). Furthermore, we observe that this mechanism enhances training stability and mitigates optimization collapse. We hypothesize that this is achieved by aligning the reward signals with the policy logits. Without this calibration, the misalignment between segmentation-related rewards and text probabilities renders the estimation of relative advantage unreliable. (3) The *Auxiliary Segmentation Loss* further enhances the discriminative capacity of the [SEG] embedding by providing direct supervision to the SAM 2 decoder. While spatial-wise and keyframe-wise losses independently improve performance, their joint application achieves optimal results (60.51% \rightarrow 61.95% on MeViS), demonstrating their complementary roles in stabilizing multi-turn RL training.

5. Conclusion

In this work, we presented **VideoSEG-O3**, a novel multi-turn reinforcement learning framework that fundamentally redefines the paradigm of Reasoning Video Object Segmentation (RVOS) by transitioning from passive, single-step inference to *active, iterative exploration*. By mimicking human cognitive strategies, our *Multi-turn Temporal-Spatial Chain-of-Thought* empowers the model to autonomously navigate complex video content through iterative spatio-temporal refocusing. This hierarchical and recursive mechanism ensures that the model maintains a holistic understanding of the video while precisely capturing the target object across dynamic scenes. Furthermore, the introduction of *SEG-aware Logit Calibration* effectively bridges the optimization gap between discrete textual reasoning and dense pixel-level prediction, enabling synchronized and efficient end-to-end optimization within the RL stage.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Liu, L., Zhang, Z., and Shou, M. Z. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 2024.
- Baidu-ERNIE-Team. Ernie 4.5 technical report, 2025.
- Chen, W., Chen, L., and Wu, Y. An efficient and effective transformer decoder-based framework for multi-task visual grounding. In *ECCV*, pp. 125–141. Springer, 2024.
- Cheng, H. K. and Schwing, A. G. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022.
- Cuttano, C., Trivigno, G., Rosi, G., Masone, C., and Averta, G. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *CVPR*, pp. 3395–3405, 2025.
- Dai, M., Cheng, W., Liu, J.-J., Yang, L., Feng, Z., Yang, W., and Wang, J. Improving generalized visual grounding with instance-aware joint learning. *PAMI*, 2025a.
- Dai, M., Cheng, W., Zhuang, J., Liu, J.-j., Zhao, H., Feng, Z., and Yang, W. Propvg: End-to-end proposal-driven visual grounding with multi-granularity discrimination. *ICCV*, 2025b.
- Dai, M., Yang, S., Duan, B., Yang, W., and Wang, J. Momentseg: Moment-centric sampling for enhanced video pixel understanding, 2025c.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- Deng, A., Chen, T., Yu, S., Yang, T., Spencer, L., Tian, Y., Mian, A. S., Bansal, M., and Chen, C. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *CVPR*, pp. 8625–8636, 2025.
- Ding, H., Liu, C., He, S., Jiang, X., and Loy, C. C. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023.
- Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.
- Gong, S., Zhang, L., Zhuge, Y., Jia, X., Zhang, P., and Lu, H. Reinforcing video reasoning segmentation to think before it segments. *arXiv preprint arXiv:2508.11538*, 2025a.
- Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., and Lu, H. The devil is in temporal token: High quality video reasoning segmentation. In *CVPR*, pp. 29183–29192, 2025b.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *ICLR*, 2021.
- Huang, J., Xu, Z., Zhou, J., Liu, T., Xiao, Y., Ou, M., Ji, B., Li, X., and Yuan, K. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *arXiv preprint arXiv:2505.22596*, 2025.
- Jin, P., Takanobu, R., Zhang, W., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024.
- Kao, S.-h., Tai, Y.-W., and Tang, C.-K. Thinkvideo: High-quality reasoning video segmentation with chain of thoughts. *arXiv preprint arXiv:2505.18561*, 2025.
- Khoreva, A., Rohrbach, A., and Schiele, B. Video object segmentation with language referring expressions. In *ACCV*, pp. 123–141, 2018.
- Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.
- Li, J., Lu, W., Fei, H., Luo, M., Dai, M., Xia, M., Jin, Y., Gan, Z., Qi, D., Fu, C., et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024a.
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pp. 22195–22206, 2024b.

- 495 Liang, T., Jiang, H., Yang, Y., Tan, C., Li, S., Zheng, W.-S.,
496 and Hu, J.-F. Long-rvos: A comprehensive benchmark
497 for long-term referring video object segmentation. *arXiv*
498 *preprint arXiv:2505.12702*, 2025a.
- 499 Liang, T., Lin, K.-Y., Tan, C., Zhang, J., Zheng, W.-S., and
500 Hu, J.-F. Referdino: Referring video object segmentation
501 with visual grounding foundations. *ICCV*, 2025b.
- 503 Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., and
504 Han, S. Vila: On pre-training for visual language models.
505 In *CVPR*, 2024.
- 507 Lin, L., Yu, X., Pang, Z., and Wang, Y.-X. Glus: Global-
508 local reasoning unified into a single large language model
509 for video segmentation. In *CVPR*, pp. 8658–8667, 2025.
- 511 Liu, C., Ding, H., and Jiang, X. Gres: Generalized referring
512 expression segmentation. In *CVPR*, pp. 23592–23601,
513 2023.
- 514 Liu, Y., Ma, Z., Pu, J., Qi, Z., Wu, Y., Ying, S., and Chen,
515 C. W. Unipixel: Unified object referring and segmenta-
516 tion for pixel-level visual reasoning. In *NeurIPS*, 2025a.
- 518 Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B.,
519 and Jia, J. Seg-zero: Reasoning-chain guided seg-
520 mentation via cognitive reinforcement. *arXiv preprint*
521 *arXiv:2503.06520*, 2025b.
- 522 Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., and
523 Jia, J. Visionreasoner: Unified visual perception and
524 reasoning via reinforcement learning. *arXiv preprint*
525 *arXiv:2505.12081*, 2025c.
- 527 Lu, Z., Li, L., Wang, J., Feng, Y., Chen, B., Chen, K., and
528 Wang, Y. Coprs: Learning positional prior from chain-
529 of-thought for reasoning segmentation. *arXiv preprint*
530 *arXiv:2510.11173*, 2025.
- 532 Miao, B., Bennamoun, M., Gao, Y., and Mian, A. Spectrum-
533 guided multi-granularity referring video object segmenta-
534 tion. In *ICCV*, pp. 920–930, 2023.
- 536 Miao, B., Bennamoun, M., Gao, Y., Shah, M., and Mian, A.
537 Temporally consistent referring video object segmenta-
538 tion with hybrid memory. *TCSVT*, 34(11):11373–11385,
539 2024.
- 541 Munasinghe, S., Thushara, R., Maaz, M., Rasheed, H. A.,
542 Khan, S., Shah, M., and Khan, F. Pg-video-llava: Pixel
543 grounding large video-language models. *arXiv preprint*
544 *arXiv:2311.13435*, 2023.
- 545 Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan,
546 F. S., and Khan, S. Videoglamm: A large multimodal
547 model for pixel-level visual grounding in videos. *arXiv*
548 *preprint arXiv:2411.04923*, 2024.
- 549 Nagaraja, V. K., Morariu, V. I., and Davis, L. S. Modeling
context between objects for referring expression under-
standing. In *ECCV*, pp. 792–807. Springer, 2016.
- Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S.,
Cholakkal, H., Anwer, R. M., Xing, E., Yang, M.-H., and
Khan, F. S. Glamm: Pixel grounding large multimodal
model. In *CVPR*, 2024.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma,
T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L.,
Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y.,
Girshick, R., Dollár, P., and Feichtenhofer, C. Sam 2:
Segment anything in images and videos. *arXiv preprint*
arXiv:2408.00714, 2024.
- Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., and
Jin, X. Pixellm: Pixel reasoning with large multimodal
model. In *CVPR*, 2024.
- Seo, S., Lee, J.-Y., and Han, B. Urvos: Unified referring
video object segmentation network with a large-scale
benchmark. In *ECCV*, 2020.
- Wang, S., Fang, G., Kong, L., Li, X., Xu, J., Yang, S., Li, Q.,
Zhu, J., and Wang, X. Pixelthink: Towards efficient chain-
of-pixel reasoning. *arXiv preprint arXiv:2505.23727*,
2025.
- Wei, C., Tan, H., Zhong, Y., Yang, Y., and Ma, L.
Lasagna: Language-based segmentation assistant for
complex queries. *arXiv preprint arXiv:2404.08506*, 2024.
- Wei, C., Zhong, Y., Tan, H., Liu, Y., Zhao, Z., Hu, J., and
Yang, Y. Hyperseg: Towards universal visual segmenta-
tion with large language model. *CVPR*, 2025a.
- Wei, C., Zhong, Y., Tan, H., Zeng, Y., Liu, Y., Zhao, Z.,
and Yang, Y. Instructseg: Unifying instructed visual
segmentation with multi-modal large language models.
ICCV, 2025b.
- Wu, J., Jiang, Y., Sun, P., Yuan, Z., and Luo, P. Language as
queries for referring video object segmentation. In *CVPR*,
2022.
- Wu, J., Zhong, M., Xing, S., Lai, Z., Liu, Z., Chen, Z.,
Wang, W., Zhu, X., Lu, L., Lu, T., et al. Visionllm
v2: An end-to-end generalist multimodal large language
model for hundreds of vision-language tasks. *NeurIPS*,
37:69925–69975, 2024.
- Xu, Z., Guo, Y., Lu, Y., Yang, F., and Li, J. Videoseg-r1:
Reasoning video object segmentation via reinforcement
learning. *arXiv preprint arXiv:2511.16077*, 2025.
- Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G.,
Xie, W., and Gavves, E. Visa: Reasoning video object
segmentation via large language models. *ECCV*, 2024.

- 550 Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L.
551 Modeling context in referring expressions. In *ECCV*, pp.
552 69–85. Springer, 2016.
- 553 Yu, S., Cho, J., Yadav, P., and Bansal, M. Self-chained
554 image-language model for video localization and question
555 answering. *NeurIPS*, 36:76749–76771, 2023.
- 557 Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong,
558 Y., Qi, L., Feng, J., and Yang, M.-H. Sa2va: Marrying
559 sam2 with llava for dense grounded understanding of
560 images and videos. *arXiv preprint arXiv:2501.04001*,
561 2025a.
- 563 Yuan, H., Sun, Y., Li, Y., Zhang, T., Deng, X., Ding, H.,
564 Qi, L., Wang, A., Li, X., and Yang, M.-H. Visual reason-
565 ing tracer: Object-level grounded reasoning benchmark.
566 *arXiv preprint arXiv:2512.05091*, 2025b.
- 567 Zhang, J., Wang, T., Ge, Y., Ge, Y., Li, X., Shan, Y.,
568 and Wang, L. Timelens: Rethinking video tempo-
569 ral grounding with multimodal llms. *arXiv preprint*
570 *arXiv:2512.14698*, 2025.
- 572 Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Chen,
573 C. L., and Yan, S. Omg-llava: Bridging image-level,
574 object-level, pixel-level reasoning and understanding. In
575 *NeurIPS*, 2024.
- 577 Zheng, R., Qi, L., Chen, X., Wang, Y., Wang, K., Qiao, Y.,
578 and Zhao, H. Villa: Video reasoning segmentation with
579 large language model. *arXiv preprint arXiv:2407.14500*,
580 2024.
- 581 Zhong, H., Zhu, M., Du, Z., Huang, Z., Zhao, C., Liu, M.,
582 Wang, W., Chen, H., and Shen, C. Omni-r1: Reinforce-
583 ment learning for omnimodal reasoning via two-system
584 collaboration. *arXiv preprint arXiv:2505.20256*, 2025.
- 586 Zhu, J., Cheng, Z.-Q., He, J.-Y., Li, C., Luo, B., Lu, H.,
587 Geng, Y., and Xie, X. Tracking with human-intent rea-
588 soning. *arXiv preprint arXiv:2312.17448*, 2023.
- 590 Zhu, L., Ouyang, B., Zhang, Y., Cheng, T., Hu, R., Shen, H.,
591 Ran, L., Chen, X., Yu, L., Liu, W., and Wang, X. Lens:
592 Learning to segment anything with unified reinforced
593 reasoning, 2025.
- 594
595
596
597
598
599
600
601
602
603
604

Appendix Overview

This appendix provides supplementary details supporting the main manuscript. The content is organized as follows:

Appendix Overview

A. Implementation Details	· A.1 Datasets and Evaluation Metrics · A.2 Training Details (SFT, CoT, RL) · A.3 Inference Strategy
B. VTS-CoT Construction Details	· B.1 Pipeline Overview · B.2 Prompt Designs
C. Additional Methods	· C.1 VideoSEG-O3 Reward Design Details
D. Additional Quantitative Results	· D.1 Results on Image Segmentation Tasks · D.2 Results on GroundMoRe Tasks
E. Additional Ablation Studies	· E.1 Analysis of RL Training Dynamics · E.2 Effect of Post-Processing Strategies · E.3 Effectiveness of Joint Segmentation Loss · E.4 Effectiveness of Reward Design
F. Additional Qualitative Results	· F.1 Multi-turn Reasoning Visualizations

A. Implementation Details

A.1. Datasets and Evaluation Metrics

Datasets. We train VideoSEG-O3 using a hierarchical three-stage pipeline, progressively endowing the model with distinct capabilities through tailored datasets. The specific composition of datasets for each stage is detailed in Table 5.

- **Stage I: Supervised Fine-Tuning (SFT).**

- *Objective:* To equip the model with core competencies—including Video QA, Image/Video Segmentation, and Temporal Index Understanding—while specifically aligning the latent representation of the [SEG] token with the SAM decoder. Deviating from conventional timestamp-based methods, we explicitly employ frame-level **temporal indices** for temporal grounding. This design serves as a crucial foundation for Stage II and III, empowering the model to navigate frame sequences and develop a robust understanding of discrete temporal relationships.

- *Datasets:*

- * **Video QA:** ChatUniVi (Jin et al., 2024) (100K).
- * **Temporal Grounding:** TimeLens-100K (Zhang et al., 2025) (100K), utilized to enforce fine-grained frame index understanding.
- * **Image RES:** RefCOCO+/g (Yu et al., 2016; Nagaraja et al., 2016) (56K) and GCG (Rasheed et al., 2024) (214K).
- * **Video RES:** Ref-YTVOS (Seo et al., 2020) (3.5K), ReVOS (Yan et al., 2024) (1.7K), MeViS (Ding et al., 2023) (0.6K), Ref-SAV (Yuan et al., 2025a) (37K), and Long-RVOS (Liang et al., 2025a) (2.2K).

- **Stage II: Chain-of-Thought Cold-Start.**

- *Objective:* To elicit multi-step reasoning, enable tool usage, and standardize the format for multi-turn interactions.
- *Datasets (VTS-CoT):* We constructed a proprietary dataset (**VTS-CoT**, 6K samples) via GPT-assisted labeling. Source clips were selected from three challenging benchmarks (ReVOS (Yan et al., 2024), MeViS (Ding et al., 2023), Long-RVOS (Liang et al., 2025a)) to ensure diversity in reasoning complexity, motion expression, and temporal duration.

660 • **Stage III: Reinforcement Learning (RL).**

- 661 – *Objective:* To optimize the model for multi-turn interactions and enhance segmentation accuracy by guiding the
662 exploration of discriminative information via reward mechanisms.
663 – *Datasets:* Training is conducted on challenging scenarios from ReVOS, and MeViS.
664

665 **Evaluation Datasets and Metrics.** We evaluate **VideoSEG-O3** across eight major benchmarks to verify its reasoning
666 and generalization capabilities. For Referring Video Object Segmentation, we report results on **MeViS** (Ding et al., 2023),
667 **Ref-Youtube-VOS** (Seo et al., 2020), **Ref-DAVIS17** (Khoreva et al., 2018), **Ref-SAV** (Yuan et al., 2025a), and **Long-**
668 **RVOS** (Liang et al., 2025a). For Reasoning Video Object Segmentation, we conduct evaluations on **ReVOS** (Yan et al.,
669 2024) (In-Domain), alongside zero-shot benchmarks on **ReasonVOS** (Bai et al., 2024) and **GroundMoRe** (Deng et al.,
670 2025) (Out-Domain). Performance is measured using standard metrics: \mathcal{J} (average Intersection over Union, IoU), \mathcal{F}
671 (boundary F-measure), and $\mathcal{J}\&\mathcal{F}$ (the average of \mathcal{J} and \mathcal{F}).
672

673 *Table 5.* The three-stage training pipeline of VideoSEG-O3, detailing the specific capabilities developed and the datasets utilized at each
674 stage.

Stage	Capability	Training Datasets
Stage I: SFT	Video QA Image/Video SEG Temporal Indices Understanding	Video QA: ChatUniVi (Jin et al., 2024) (100K) Temporal Grounding: TimeLens-100K (Zhang et al., 2025) (100K) Image RES: RefCOCO+/g (Yu et al., 2016) (56K), ReasonSeg (0.2K), GCG (214K) Video RES: Ref-YTVOS (Seo et al., 2020) (3.5K), ReVOS (Yan et al., 2024) (1.7K), MeViS (Ding et al., 2023) (0.6K), Ref-SAV (Yuan et al., 2025a) (37K), Long-RVOS (Liang et al., 2025a) (2.2K)
Stage II: CoT Cold-Start	Step-wise Reasoning Tool Usage	VTS-CoT (6K, Ours): Curated from ReVOS (Yan et al., 2024), Long-RVOS (Liang et al., 2025a), and MeViS (Ding et al., 2023) via GPT-assisted labeling.
Stage III: GRPO-RL	Multi-turn Interaction Mask Refinement	Video RES: MeViS (Ding et al., 2023) (0.6K), ReVOS (Yan et al., 2024) (1.7K)

690
691 **A.2. Training Details**

692
693 The training of VideoSEG-O3 is hierarchically structured into three stages: supervised fine-tuning (SFT), Chain-of-Thought
694 (CoT) cold-start, and Reinforcement Learning (RL). This section provides the exhaustive hyperparameter configurations for
695 each phase.
696

697 **A.2.1. STAGE I (SFT) AND STAGE II (CoT COLD-START)**

698 The transition from Stage I to Stage II marks a significant shift from static image-level alignment to dynamic multi-frame
699 temporal reasoning. While Stage I focuses on the fundamental modality bridge using static samples, Stage II introduces a
700 decoupled visual sampling strategy to support the multi-turn Chain-of-Thought process. The detailed hyperparameter and
701 visual input configurations are summarized in Table 6.
702

703 **Decoupled Visual Input.** A pivotal difference in Stage II is the hierarchical processing of visual information. Specifically,
704 we introduce up to 20 **temporal frames** sampled at a low resolution ($32 \times 28 \times 28$ max pixels) to provide global motion
705 context without exceeding the MLLM’s token limit. Simultaneously, a set of **spatial frames** (up to 5) is provided at a
706 moderate resolution ($128 \times 28 \times 28$ max pixels) to capture environmental details. Crucially, for each reasoning turn, a
707 single **keyframe** is injected at the highest resolution ($512 \times 28 \times 28$ max pixels).
708

709 **A.2.2. STAGE III: REINFORCEMENT LEARNING (GRPO)**

710 Stage III leverages Group Relative Policy Optimization (GRPO) (DeepSeek-AI, 2025) to facilitate autonomous exploration
711 and refine the model’s decision-making policy. Unlike previous stages, the RL phase employs full-parameter fine-tuning
712 (excluding the frozen visual encoder) to maximize the expressive capacity of the policy. The detailed hyperparameter and
713 visual configurations are summarized in Table 7.
714

Table 6. Experimental settings and visual input configurations for Stage I (SFT) and Stage II (CoT Cold-start).

Configuration	Stage I (SFT)	Stage II (CoT Cold-start)
LoRA Rank (r)	128	128
LoRA Alpha (α)	256	256
Backbone Frozen	Visual Encoder, LLM	Visual Encoder, LLM
Trainable Modules	SAM2 Decoder	SAM2 Decoder
<i>Visual Sampling Strategy</i>		
Temporal Frames (Max)	None	20
Temporal Max Pixels	N/A	$32 \times 28 \times 28$
Spatial Frames (Max)	1	5
Spatial Max Pixels	$512 \times 28 \times 28$	$128 \times 28 \times 28$
Keyframe per Round	N/A	1
Keyframe Max Pixels	N/A	$512 \times 28 \times 28$
Peak Learning Rate	4×10^{-5}	4×10^{-5}
Warmup Ratio	0.05	0.05
Max Sequence Length	8192	8192

Dynamic Visual State in RL. During the RL stage, the visual observation space \mathcal{V}_t is updated dynamically to balance reasoning depth with computational efficiency. While spatial frame and keyframe resolutions remain consistent with the CoT stage, the temporal window is more strictly managed. We limit the total temporal frames to 10, with up to 5 newly sampled temporal frames added in each interaction round based on exploration actions. This ensures the model learns to prioritize informative temporal intervals within a compact context window.

A.3. Inference Strategy

The inference phase maintains the multi-turn reasoning logic established in the RL stage but employs enhanced visual sampling strategies to optimize zero-shot transfer performance. Detailed configuration is provided in Table 8, with key technical enhancements highlighted below:

- **Spatial Sampling Density:** To improve environment perception, we increase the number of **spatial frames** to 8, compared to the 5 frames used during RL training.
- **Resolution Scaling:** While the temporal and keyframe resolutions remain consistent with Stage III, the **spatial resolution** is enhanced to $256 \times 28 \times 28$ pixels (doubling the training resolution) to provide clearer semantic cues for object localization.
- **Execution Constraints:** The system is constrained to a maximum of 3 interaction turns, processing up to 20 sampled video frames to ensure a comprehensive temporal horizon.

Table 7. Reinforcement Learning (Stage III) experimental settings. This stage prioritizes efficient multi-turn exploration through constrained temporal windows and high-resolution spatial anchors.

RL Configuration	Value
Algorithm	GRPO
Trainable Modules	LLM, SAM2 Decoder
Optimization Policy	Full-parameter Fine-tuning
Learning Rate	3×10^{-6}
KL Coefficient (β)	0.04
Number of Generations (G)	4
Max Training Steps	1000
Max Interaction Turns	3
<i>Visual Sampling (RL-specific)</i>	
Initial Temporal Frames (Max)	10
New Temporal Frames per Round (Max)	5
Spatial Frames (Max)	5
<i>Resolution Constraints</i>	
Temporal Max Pixels	$32 \times 28 \times 28$
Spatial Max Pixels	$128 \times 28 \times 28$
Keyframe Max Pixels	$512 \times 28 \times 28$
Auxiliary Loss Ratio (λ_{seg})	0.2
Numerical Precision	bfloat16

Table 8. Inference hyperparameter configurations for VideoSEG-O3. Values in parentheses denote the increments or scaling factors relative to the Reinforcement Learning training stage (Stage III).

Inference Parameter	Value
Max Interaction Turns	3
Max Video Sample Frames	20
Spatial Sampling Frames (K)	8 (+3 vs. RL)
<i>Resolution Constraints (Max Pixels)</i>	
Temporal Resolution	$32 \times 28 \times 28$
Spatial Resolution	$256 \times 28 \times 28$ ($\times 2$ vs. RL)
Keyframe Resolution	$512 \times 28 \times 28$

B. VTS-CoT Construction Details

B.1. Pipeline Overview

As illustrated in Fig. 6, the construction pipeline of VTS-CoT is structured into four distinct components: data curation, temporal labeling, candidate generation, and Chain-of-Thought (CoT) synthesis. In this section, we elaborate on the design rationale and implementation details for each component.

1. Diverse Data Curation. We aggregate a diverse collection of video clips from three representative benchmarks: ReVOS, Long-RVOS, and MeViS. Each dataset introduces unique challenges to ensure the complexity and generalization capability of VTS-CoT. Specifically, ReVOS focuses on reasoning-intensive scenarios, MeViS emphasizes complex motion and expression understanding, while Long-RVOS contributes samples characterized by extended temporal durations.

2. Temporal Labeling and Mask Quality Assessment. We design a specialized prompt (see Table 9) to guide the Qwen3VL-235B-A30B-Instruct (Bai et al., 2025) model in performing precise temporal labeling and evaluating mask fidelity. Following the initial annotation, we implement a rigorous filtering protocol to retain high-quality samples. To ensure annotation accuracy, we employ a logical dual-verification mechanism that cross-references the temporal grounding results with the ground-truth mask targets; samples exhibiting significant discrepancies are discarded. Furthermore, to enhance the model’s temporal grounding precision, we adopt a high-resolution, dense sampling strategy and explicitly inject mask information into the prompt to serve as visual auxiliary cues.

3. Temporal Candidation. Leveraging the refined temporal annotations from the previous stage, we generate three top-ranked candidate temporal intervals. This process is executed by the ERNIE-45-VL-424B-A47B (Baidu-ERNIE-Team, 2025) model using the prompt detailed in Table 10. To encourage the generation of diverse yet plausible temporal proposals, we deliberately employ a low-resolution, sparse video sampling strategy during this phase.

4. Decoupled CoT Construction. With the high-quality temporal annotations and candidate intervals established, we utilize the prompt presented in Table 11 to instruct the Qwen3VL-235B-A30B-Thinking (Bai et al., 2025) model in synthesizing the final reasoning chains. To maximize the diversity and coverage of the generated Chain-of-Thought (CoT) data, we apply a hybrid video sampling strategy, alternating between high-resolution dense sampling and low-resolution sparse sampling to simulate varying observational contexts.

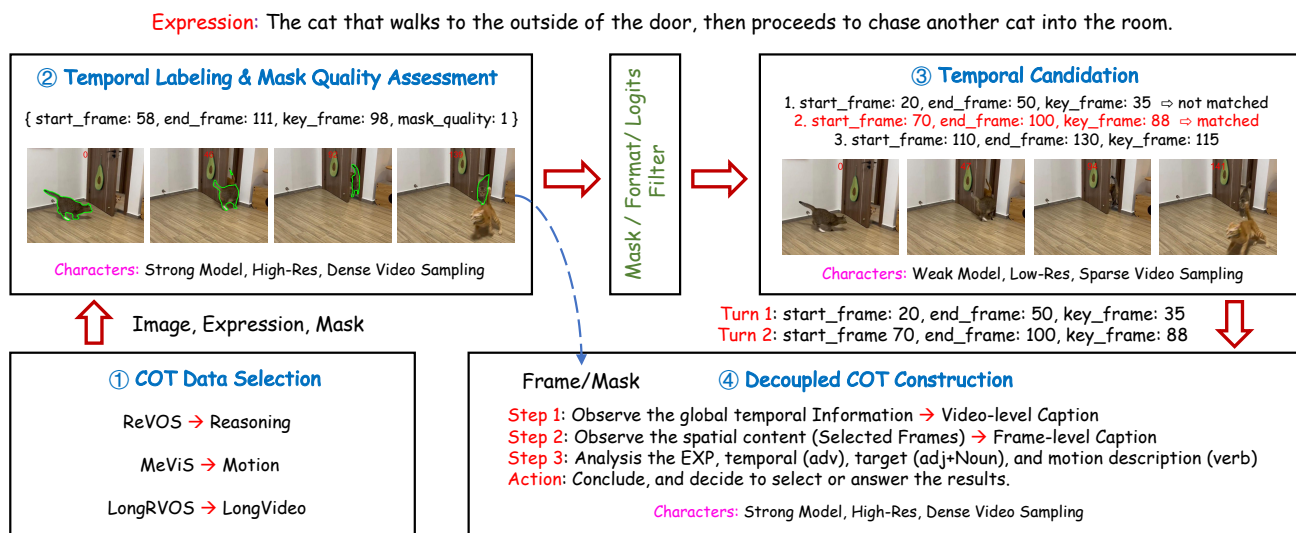


Figure 6. Pipeline of our VTS-CoT: Data selection → Temporal labeling → Candidate generation → Chain-of-Thought construction.

B.2. Prompt Designs

We provide the prompt templates used for the distinct stages of VTS-CoT construction below.

Prompt for Temporal Labeling and Mask Quality Assessment

Role & Objective:

Act as a specialist in Spatiotemporal Video Grounding and Annotation Verification. Your objective is to precisely localize a specific event temporally and evaluate the fidelity of pre-annotated spatial segmentation masks.

Input Data:

- **Target Query:** <referring_expression>
- **Sampled Frames:** A uniform sequence of video frames containing:
 - *Visual Indices:* **Red numbers** indicating the temporal index of each frame.
 - *Candidate Masks:* **Green contour** outlines representing the ground truth proposals.
- **Video Metadata:** Total frame count:<total_frames>, denoted as T_{total} .

Tasks:

1. **Temporal Localization:** Identify the most semantic-relevant time interval $[t_{start}, t_{end}]$ and a representative keyframe t_{key} that best aligns with the *Target Query*.
2. **Mask Fidelity Assessment:** Evaluate the *Candidate Masks*. Determine if the green contours accurately and consistently capture the target object described in the query.

Constraints & Rules:

- **Temporal Duration:** The selected window must capture the core event without being excessive. The duration $\Delta t = t_{end} - t_{start}$ must satisfy:

$$5 < \Delta t < 0.5 \times T_{total}$$

- **Index Validity:** Ensure logical consistency of frame indices:

$$0 \leq t_{start} \leq t_{key} \leq t_{end} < T_{total}$$

- **Quality Criteria:**

- **1 (Valid):** The mask is complete, temporally consistent, and semantically correct.
- **0 (Invalid):** The mask is fragmented, drifting, absent, or misaligned with the target.

Output Schema:

Return strictly a single JSON object with the following structure:

```
{"start_frame": <int>, "end_frame": <int>, "key_frame": <int>, "mask_quality": <0|1>}
```

Table 9. The structured prompt template used for GPT-assisted dataset curation. We formalize the constraints using mathematical inequalities to enforce precise temporal localization.

Prompt for Top-3 Non-Overlapping Temporal Candidation

Role & Objective:

Act as a specialist in Multi-Event Video Analysis. Your task is to identify and rank the Top-3 most relevant, non-overlapping temporal segments within a video that align with a specific text query.

Input Data:

- **Target Query:** <referring-expression>
- **Visual Context:** Sampled video frames with visual indices (red numbers) indicating temporal location.
- **Metadata:** Total frame count:<total_frames>, denoted as T_{total} .

Iterative Exclusion Strategy:

To ensure diversity and strict ranking, follow this sequential selection process:

1. **Primary Selection (S_1):** Identify the single most relevant segment in the entire video.
2. **Secondary Selection (S_2):** Mask the time range of S_1 . Find the best matching segment in the remaining video parts ($V \setminus S_1$).
3. **Tertiary Selection (S_3):** Mask the time ranges of both S_1 and S_2 . Find the best match in ($V \setminus (S_1 \cup S_2)$).

Constraints & Rules:

- **Relevance Ranking:** Output must be sorted by semantic relevance: $Rel(S_1) > Rel(S_2) > Rel(S_3)$.
- **Non-Overlapping:** Segments must be mutually exclusive:

$$S_i \cap S_j = \emptyset, \quad \forall i \neq j$$

- **Duration Constraints:** For any segment S_i with duration Δt_i :

$$0.1 \times T_{total} \leq \Delta t_i < 0.33 \times T_{total}$$

- **Index Validity:** $0 \leq t_{start} \leq t_{key} \leq t_{end} < T_{total}$.
- **Exception Handling:** Return fewer than 3 segments only if the video lacks sufficient distinct content.

Output Schema:

Output **ONLY** a raw JSON list. Do not use markdown blocks.

```
[ {"start_frame": <int>, "end_frame": <int>, "key_frame": <int>},
  {"start_frame": <int>, "end_frame": <int>, "key_frame": <int>, ... } ]
```

Table 10. The prompt template for the Top-3 iterative temporal grounding task. We explicitly enforce an exclusion strategy to generate diverse, non-overlapping proposals.

Prompt for Sequential Spatiotemporal CoT Construction

Role & Objective:

Act as an Advanced Spatiotemporal Reasoning Agent. Simulate a strictly sequential human reasoning process to locate a target described by a text expression. Generate a logical "Chain of Thought" (CoT) justifying segment selection, moving from global overview to specific verification.

Task Simulation (Two Phases):

- Phase 1 (Initialization):** Analyze global context and text to form a hypothesis. **Constraint:** You cannot access specific segment details yet. **Action:** Request Segment 0.
- Phase 2 (Verification Loop):** Analyze specific frames requested in the previous step. Verify if the target matches. **Action:** Request next segment OR Confirm final target.

Input Data:

- **Global:** Referring Expression, Total Video Length, Sampled Frame Indices.
- **Input Segments:** Sequence of temporal segments (start/end/keyframe).
- **Visual Cues:** **Red Numbers** (Frame Indices), **Green Contours** (Target location, internal use only).

Strict Reasoning Workflow (Output JSON list of size $N + 1$):

- First Item (Global Analysis):**
 - *Step 1 (Global Temporal):* Analyze low-res global context for candidates.
 - *Step 2 (Local Spatial):* Inspect Sampled Frames.
 - *Step 3 (Alignment):* Correlate visual cues with text constraints.
 - *Action:* Request Input Segment [0].
- Subsequent Items (Verification Loop for Segment k):**
 - *Step 1 (Segment Temporal):* Analyze motion in the requested interval.
 - *Step 2 (Segment Spatial):* Verify visual attributes in the requested `key_frame`.
 - *Step 3 (Refinement):* Compare specific evidence against text.
 - *Action:* Justify checking Segment [k+1] OR Confirm Final Segment.

Constraints & Rules:

- **Information Isolation:** You only see data requested in the *previous* Action. No peeking ahead.
- **No Visual Leakage:** Never mention "green contours" or "outlines" in the output. Describe the object itself.
- **Format:** Single-line, valid raw JSON list. No markdown.

Few-Shot Strategy (Example: Single Segment):

Exp: "The car at the very front at the beginning..."

```
[ {"think": "Step 1: Global view shows a race... Step 2: At frame 0, red car leads... Step 3: 'beginning' requires focusing on leader. Action: Inspect 0-23 segment...", "start_frame": 0, "end_frame": 23, "key_frame": 5}, {"think": "Step 1: Segment 0-23 shows yellow Mercedes... Step 2: Keyframe 4 shows it at front... Step 3: Matches 'very front'. Action: Target confirmed.", "start_frame": 0, "end_frame": 10, "key_frame": 4} ]
```

Output Schema:

Output **ONLY** a raw JSON list.

```
[ {"think": "Step 1: [Temporal]. Step 2: [Spatial]. Step 3: [Expression]. Action: [Decision].", "start_frame": <int>, "end_frame": <int>, "key_frame": <int>}, ... ]
```

Table 11. The system prompt for the Spatiotemporal Reasoning Agent. It enforces a strict, two-phase CoT workflow (Global Analysis → Sequential Verification) with information isolation to simulate human-like video grounding.

C. Additional Methods

C.1. VideoSEG-O3 Reward Design Details

In this section, we provide the mathematical formulation for the composite reward function \mathcal{R} , which consists of four components: Format Reward (\mathcal{R}_f), Temporal Reward (\mathcal{R}_t), Segmentation Quality Reward (\mathcal{R}_m), and Progressive Reward (\mathcal{R}_p).

C.1.1. FORMAT REWARD (\mathcal{R}_f)

To ensure the model adheres to the strict protocols required for parsing and execution, we implement a two-stage format check.

- **Think Format:** We parse the `<think>` block to verify the existence of the Chain-of-Thought structure. The reward is accumulated based on the presence of four required keys: “Step 1”, “Step 2”, “Step 3”, and “Action”. Each successfully detected key yields +0.25, summing to a maximum of 1.0.
- **Result Format:** We validate the syntactic correctness of the final answer block. The output must strictly follow the format: a `<VTG>` dictionary containing `start`, `end`, and `keyframe` fields, followed by a `<RefVOS>` tag enclosing the `[SEG]` token. If the output cannot be parsed or the intervals are invalid ($start \geq end$), a penalty (e.g., -0.5) is applied; otherwise, it is eligible for subsequent dense rewards.

C.1.2. TEMPORAL REWARD (\mathcal{R}_t)

The temporal reward evaluates the model’s grounding capability from two perspectives: discrete keyframe selection and continuous temporal range generation. Let $E_m \in \{0, 1\}^T$ be the ground-truth binary sequence and $[S, E]$ be the predicted temporal interval.

Keyframe Alignment Reward (\mathcal{R}_t^k): This component validates whether the selected keyframe index k falls within the valid action range. It acts as a strict binary feedback:

$$\mathcal{R}_t^k = \begin{cases} +1.0 & \text{if } E_m[k] = 1, \\ -1.0 & \text{if } E_m[k] = 0. \end{cases} \quad (9)$$

Temporal Precision Reward (\mathcal{R}_t^p): To evaluate the quality of the predicted interval $[S, E]$, we calculate the temporal precision P_t , defined as the proportion of the predicted frames that actually contain the target: $P_t = \frac{\sum_{t=S}^E E_m[t]}{E-S}$. The dense reward is then calculated as:

$$\mathcal{R}_t^p = \begin{cases} 0.5 \times P_t & \text{if } P_t \geq 0.5, \\ -0.5 & \text{if } P_t < 0.5. \end{cases} \quad (10)$$

This formulation penalizes loose predictions (low precision) with a fixed penalty of -0.5 , while scaling the positive reward linearly with precision when it exceeds the 0.5 threshold.

C.1.3. SEGMENTATION QUALITY REWARD (\mathcal{R}_m)

To stabilize training, we map continuous IoU values to discrete reward tiers using piecewise functions.

Spatial Reward (\mathcal{R}_m^s): This rewards the average mask quality across sampled spatial frames. Let IoU_s be the average IoU of the predicted masks in the final turn. The mapping function is defined as:

$$\mathcal{R}_m^s = \begin{cases} \min(3(\text{IoU}_s - 0.4), 1.5) & \text{if } \text{IoU}_s > 0.4, \\ 0.0 & \text{if } 0.2 < \text{IoU}_s \leq 0.4, \\ -1.0 & \text{if } \text{IoU}_s \leq 0.2. \end{cases} \quad (11)$$

Keyframe Superiority Reward (\mathcal{R}_m^k): This component incentivizes the model to identify a keyframe that offers a significantly better view than the average spatial context. It is based on the differential $\Delta_k = \text{IoU}_k - \text{IoU}_s$:

$$\mathcal{R}_m^k = \begin{cases} \min(5(\Delta_k - 0.1), 2.0) & \text{if } \Delta_k > 0.2, \\ 0.5 & \text{if } 0.1 < \Delta_k \leq 0.2, \\ 0.0 & \text{if } -0.05 < \Delta_k \leq 0.1, \\ -1.0 & \text{if } \Delta_k \leq -0.05. \end{cases} \quad (12)$$

C.1.4. PROGRESSIVE REWARD (\mathcal{R}_p)

To encourage continuous improvement across multi-turn interactions, we compare the current keyframe IoU ($\text{IoU}_{k,t}$) with the best performance from previous turns ($\text{IoU}_{\text{best}} = \max_{j < t} \text{IoU}_{k,j}$). Let $\Delta_p = \text{IoU}_{k,t} - \text{IoU}_{\text{best}}$. The progressive reward is defined as:

$$\mathcal{R}_p = \begin{cases} \min(10(\Delta_p - 0.05), 2.0) & \text{if } \Delta_p > 0.1, \\ 0.5 & \text{if } 0.05 < \Delta_p \leq 0.1, \\ 0.0 & \text{if } -0.05 < \Delta_p \leq 0.05, \\ -1.0 & \text{if } \Delta_p \leq -0.05. \end{cases} \quad (13)$$

This mechanism strictly penalizes degradation ($\Delta_p \leq -0.05$) and offers substantial rewards (slope of 10) only when the model breaks previous records by a significant margin ($\Delta_p > 0.1$).

Table 12. Comparison of model results across image segmentation tasks. We compare our VideoSEG-O3 with both large-scale MLLM-based methods and lightweight specialist methods. The metric is cIoU.

Method	RefCOCO			RefCOCO+			RefCOCog		ReasonSeg (ZS)	
	val	testA	testB	val	testA	testB	val(U)	test(U)	val	test
<i>Specialist Methods</i>										
ReLA (Liu et al., 2023)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	-	-
EEVG (Chen et al., 2024)	79.5	80.9	77.4	71.9	76.7	66.3	73.6	73.5	-	-
PropVG (Dai et al., 2025b)	82.0	83.6	80.0	77.1	79.8	72.2	77.0	77.7	-	-
InstanceVG (Dai et al., 2025a)	81.4	83.1	79.3	76.6	79.5	71.6	75.9	76.6	-	-
<i>MLLM Methods</i>										
PixelLM-7B (Ren et al., 2024)	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-
LISA-7B (Lai et al., 2024)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	61.3	62.9
GLaMM-7B (Rasheed et al., 2024)	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-
VisionLLMv2-7B (Wu et al., 2024)	79.2	82.3	77.0	68.9	75.8	61.8	73.3	74.8	56.9	48.3
OMG-LLaVA-7B (Zhang et al., 2024)	75.6	77.7	71.2	65.6	69.7	58.9	70.7	70.2	-	-
LaSagnA-7B (Wei et al., 2024)	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	48.8	47.2
VISA-7B (Yan et al., 2024)	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	52.7	57.8
VRS-HQ-7B (Gong et al., 2025b)	73.5	77.5	69.5	61.7	67.6	54.3	66.7	67.5	51.8	52.9
Sa2VA-4B (Yuan et al., 2025a)	78.9	-	-	71.7	-	-	74.1	-	-	-
<i>RL-based Methods</i>										
Seg-Zero-7B (Liu et al., 2025b)	-	80.3	-	-	76.2	-	-	72.6	52.6	52.4
PixelThink-7B (Wang et al., 2025)	-	79.3	-	-	74.8	-	-	73.9	62.7	55.8
VideoSeg-R1-7B (Xu et al., 2025)	78.2	82.3	75.1	71.8	76.1	64.7	73.1	74.1	68.2	-
VideoSEG-O3-4B (Ours)	80.9	82.5	80.5	75.5	79.9	70.6	78.9	79.0	67.7	61.5

D. Additional Quantitative Results

D.1. Results on Image Referring/Reasoning Segmentation Tasks

Although VideoSEG-O3 is primarily designed for video scenarios, image-level segmentation data was incorporated during the Supervised Fine-Tuning (SFT) phase. To evaluate whether image-level capabilities are preserved following video-centric training, we conducted benchmarks on standard image datasets. As demonstrated in Table 12, VideoSEG-O3 maintains competitive performance on image tasks, suggesting that the model effectively generalizes across both static and temporal domains.

D.2. Results on GroundMoRe Tasks

GroundMoRe (Deng et al., 2025) is a challenging benchmark requiring joint spatio-temporal reasoning for comprehensive video understanding. To evaluate the generalization capability of VideoSEG-O3, we conducted zero-shot evaluations on this dataset. The results, summarized in Table 13, demonstrate that VideoSEG-O3 achieves a new state-of-the-art with an Overall $\mathcal{J}\&\mathcal{F}$ score of **31.96%**, significantly outperforming the previous best method, MORA, by **+8.83%**. Specifically, our model yields remarkable improvements in the *Descriptive* (+14.76%) and *Causal* (+13.06%) categories. This highlights the robustness of our Multi-turn CoT in capturing fine-grained spatio-temporal attributes and tracking complex cause-effect chains. However, we observe a slight performance dip in the *Counterfactual* category (-1.31% compared to MORA). We attribute this primarily to the training data distribution gap inherent in the zero-shot setting.

E. Additional Ablation Studies

E.1. Analysis of RL Training Dynamics

Fig. 7 illustrates the optimization progression during the reinforcement learning stage across three key dimensions. First, the **reward curve** in Fig. 7(a) exhibits a steady upward trajectory, confirming the effective convergence of the policy toward higher segmentation fidelity. Second, Fig. 7(b) shows that the **average interaction turns** experience a sharp initial decline. This indicates that the RL process successfully rectifies the rigid patterns inherited from the CoT cold-start phase by eliminating redundant or irrelevant turns. Subsequently, the model enters an exploratory phase where turns slightly increase

Table 13. Comparison of Motion-Grounded Video Reasoning results on our benchmark in a zero-shot setting. Best results are highlighted in bold.

Methods	Overall			Causal			Sequential			Counterfactual			Descriptive		
	J&F	J	F	J&F	J	F	J&F	J	F	J&F	J	F	J&F	J	F
Random Baseline															
Title+ReferFormer (Wu et al., 2022)	10.13	9.73	10.53	10.50	9.91	11.08	9.06	8.56	9.57	9.42	9.07	9.78	11.38	11.21	11.55
RVOS Baseline															
ReferFormer (Wu et al., 2022)	12.72	11.15	14.29	12.67	10.95	14.38	10.73	9.35	12.10	12.36	11.12	13.61	15.15	13.26	17.04
SgMg (Miao et al., 2023)	17.49	15.81	19.16	18.94	17.04	20.85	15.46	13.97	16.96	16.35	15.24	17.45	18.77	16.69	20.86
HTR (Miao et al., 2024)	14.48	12.71	16.24	15.66	13.82	17.51	12.46	10.78	14.14	13.51	12.16	14.86	15.94	13.80	18.08
LMPM (Ding et al., 2023)	13.34	12.71	13.97	13.71	13.13	14.30	11.30	10.60	12.01	13.16	12.64	13.68	14.33	13.60	15.05
Image Reasoning Segmentation Baseline															
LISA-7B (Lai et al., 2024)	6.11	4.89	7.32	5.53	4.38	6.69	5.19	4.07	6.32	5.88	4.77	6.98	7.56	6.11	9.00
PixelLM-7B (Ren et al., 2024)	9.96	9.93	9.99	9.52	9.39	9.65	9.21	9.15	9.27	9.82	9.76	9.87	10.94	11.02	10.85
Video Reasoning Segmentation Baseline															
PG-Video-LLaVA (Munasinghe et al., 2023)	11.17	11.47	10.89	10.57	10.82	10.33	10.93	11.15	10.71	10.36	10.37	10.34	12.51	13.04	11.99
PG-Video-LLaVA+SAM2 (Ravi et al., 2024)	11.85	11.15	12.55	13.35	12.84	13.85	3.85	2.49	5.22	15.52	15.16	15.89	12.11	11.37	12.85
VISA (Yan et al., 2024)	5.31	4.72	5.90	6.07	5.50	6.64	4.48	3.80	5.15	5.10	4.34	5.86	5.39	5.02	5.77
Two-Stage Baseline															
ViLA (Lin et al., 2024)+ReferFormer	16.37	14.84	17.89	15.92	14.33	17.52	16.30	14.94	17.67	13.63	12.35	14.91	19.61	17.78	21.45
ViLA (Lin et al., 2024)+SgMg	17.07	15.25	18.89	17.34	15.43	19.26	16.30	14.57	18.04	15.31	13.96	16.67	19.18	16.96	21.40
ViLA (Lin et al., 2024)+HTR	15.72	13.90	17.53	16.46	14.60	18.31	15.11	13.39	16.83	12.61	11.22	13.99	18.37	16.10	20.63
VideoChat2 (Li et al., 2024b)+ReferFormer	15.38	13.84	16.93	14.74	13.10	16.37	13.95	12.69	15.21	13.12	11.80	14.43	19.82	17.89	21.74
VideoChat2 (Li et al., 2024b)+SgMg	16.65	14.83	18.48	16.64	14.75	18.54	14.70	13.02	16.38	15.22	13.87	16.58	20.01	17.68	22.34
VideoChat2 (Li et al., 2024b)+HTR	15.08	13.25	16.90	15.91	14.00	17.81	13.91	12.33	15.50	11.97	10.56	13.37	18.17	15.82	20.52
SeViLA (Yu et al., 2023)+ReferFormer	20.37	18.99	21.75	21.29	19.86	22.72	18.27	17.06	19.48	17.56	16.37	18.75	24.01	22.35	25.66
SeViLA (Yu et al., 2023)+SgMg	22.34	20.92	23.75	21.60	20.17	23.02	21.68	20.47	22.89	20.46	19.32	21.61	24.53	22.83	26.22
SeViLA (Yu et al., 2023)+HTR	21.75	19.85	23.64	22.28	20.42	24.15	20.37	18.60	22.14	19.70	18.03	21.36	24.41	22.11	26.71
MORA (Deng et al., 2025)	23.13	22.79	23.46	21.62	21.46	21.79	22.43	22.03	22.83	26.37	25.87	26.88	22.63	22.30	22.97
VideoSEG-O3-2B (Ours)	29.06	27.39	30.72	30.93	29.21	32.65	26.97	25.42	28.52	22.04	20.62	23.46	35.49	33.54	37.44
VideoSEG-O3-4B (Ours)	31.96	30.19	33.73	34.68	32.88	36.47	29.71	28.14	31.28	25.06	23.51	26.60	37.39	35.21	39.56

and fluctuate, eventually stabilizing as the model learns to autonomously allocate necessary reasoning rounds to maximize rewards. Finally, as shown in Fig. 7(c), the **completion length** follows a similar trend to the interaction turns, further validating that VideoSEG-O3 learns to generate more concise and semantically efficient reasoning trajectories through iterative refinement.

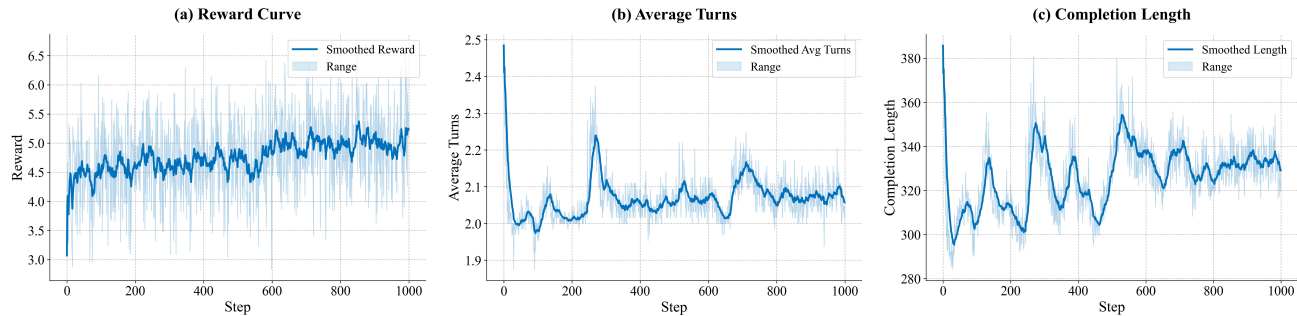


Figure 7. Reinforcement learning training curves. (a) The smoothed reward curve demonstrates consistent convergence and performance improvement. (b) The average turns indicate the model’s transition from fixed reasoning patterns to adaptive multi-turn exploration. (c) The completion length reflects the refinement and stabilization of the generated reasoning trajectories.

E.2. Effect of Post-Processing Strategies

To investigate optimal inference strategies for leveraging multi-turn reasoning in high-fidelity mask generation, we evaluate three post-processing methods: *Sa2VA-type First N*, *Keyframe Bidirectional*, and *Language Injection* (Table 14). The significant performance gains observed with the *Keyframe Bidirectional* strategy validate VideoSEG-O3’s proficiency in its two primary optimization objectives: **Keyframe Localization** and **Embedding Refinement**. By successfully identifying the most discriminative keyframe where the target is salient, the model establishes a robust visual anchor for bidirectional temporal propagation, demonstrating that precise spatiotemporal understanding is critical for maximizing segmentation performance in multi-turn interactions.

Table 14. Ablation study on post-processing strategies during inference. *Keyframe Bidirectional* significantly outperforms the baseline, highlighting the model’s strong keyframe localization ability.

Strategy	MeViS (val_u)	Long-RVOS
Sa2VA-type First N	59.95	47.22
+ Keyframe Initialization	61.54	48.15
+ [SEG] Injected Propagation	63.07	49.55

E.3. Effectiveness of Joint Segmentation Loss

To validate the effectiveness of the auxiliary segmentation loss proposed in Sec. 3.3, we conduct ablation studies by varying its weight ratio to modulate its influence during training. As depicted in Fig. 8, increasing λ_{seg} yields consistent performance gains on the in-domain MeViS dataset. Conversely, on the out-of-domain ReasonVOS dataset, we observe an initial improvement followed by a decline. Notably, this tendency becomes more pronounced as the number of training epochs increases. These results suggest that while the auxiliary loss enhances adaptation to the training distribution, an excessive weight induces overfitting, thereby compromising generalization capabilities. Consequently, we adopt $\lambda_{seg} = 0.2$ to achieve an optimal balance between in-domain performance and generalizability.

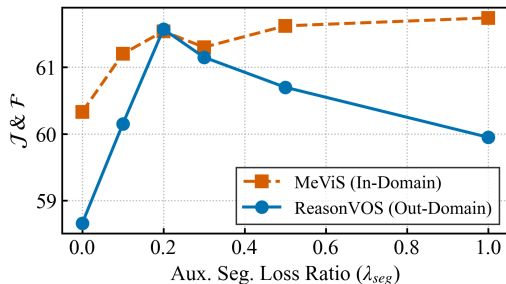


Figure 8. Ablation study on auxiliary segmentation loss (λ_{seg}).

Table 15. Ablation study on reward design. \mathcal{R}_p , \mathcal{R}_{km} , and \mathcal{R}_t denote Progressive Mask Improvement, Keyframe Mask Quality, and Temporal Accuracy rewards, respectively.

Model	MeViS (val_u)	Ref-SAV	Avg. Rounds
Step-wise Reward	57.88	50.18	1.63
Episodic Reward	61.19	52.91	2.13
w/o \mathcal{R}_p	60.37 (-0.82)	52.62 (-0.29)	2.46 (\uparrow 15.5%)
w/o \mathcal{R}_{km}	60.19 (-1.00)	50.99 (-1.92)	2.36 (\uparrow 10.8%)
w/o \mathcal{R}_t	58.70 (-2.49)	51.11 (-1.80)	2.55 (\uparrow 19.7%)

E.4. Effectiveness of Reward Design

As shown in Table 15, we investigate the impact of distinct reward components on both segmentation accuracy and interaction efficiency. The *Progressive Mask Improvement reward* (\mathcal{R}_p) is critical for regulating reasoning length; removing \mathcal{R}_p results in redundant exploration, increasing the average rounds from 2.13 to 2.46. This confirms that penalizing stagnant mask quality effectively suppresses unnecessary dialogue turns. Furthermore, the *Keyframe Mask Quality reward* (\mathcal{R}_{km}) and *Temporal Accuracy reward* (\mathcal{R}_t) drive the model toward precise spatiotemporal grounding. Notably, the omission of \mathcal{R}_t leads to the most significant performance drop on MeViS (61.19% \rightarrow 58.70%), underscoring its role in ensuring the model identifies the correct temporal interval before attempting dense segmentation.

We further explore the impact of replacing episodic rewards with step-wise rewards during training. In this setting, the reward is calculated at each step based on varying maximum interaction horizons (e.g., for a 4-turn limit, rewards are computed for horizons $T \in \{1, 2, 3, 4\}$). Our empirical analysis reveals that this step-wise approach leads to significant *reward hacking* and eventual reward collapse. Specifically, the policy tends to converge toward predicting with minimal interaction turns, failing to exploit the advantages of multi-turn reasoning. This phenomenon occurs because the model prioritizes maximizing

1320 immediate rewards within fewer steps, thereby diminishing its focus on the incremental gains between turns. Consequently,
1321 the model loses the capacity for effective multi-round refinement, leading to a degradation in overall performance.
1322

1323 **F. Additional Qualitative Results** 1324

1325 **F.1. Multi-turn Reasoning Visualizations** 1326

1327 To intuitively illustrate the multi-turn reasoning capabilities of VideoSEG-O3, we provide visualizations of representative
1328 conversational cases. We explicitly present the complete progression, including system prompts, per-round user instructions,
1329 visual inputs, chain-of-thought reasoning processes, and the final predicted masks.

1330 **Two-round Dialogues.** As shown in Fig. 9, we first demonstrate the model’s performance in two-round interaction scenarios.
1331 In these cases, VideoSEG-O3 effectively utilizes immediate user feedback to correct initial ambiguities. By integrating
1332 historical reasoning with new instructions, the model refines its understanding of the scene, allowing it to swiftly adjust its
1333 focus and achieve accurate object grounding after a single observation step.
1334

1335 **Three-round Dialogues.** Fig. 10 and Fig. 11 further showcase the model’s proficiency in handling intricate target localization
1336 tasks through a three-round iterative process. These visualizations reveal a clear coarse-to-fine refinement trajectory, where
1337 the model progressively narrows down the target by synthesizing cumulative visual evidence and multi-step feedback. From
1338 an information-theoretic perspective, this progressive interaction proves more efficient than processing exhaustive video
1339 content in a single pass, offering a robust paradigm for precise segmentation in complex dynamic environments.
1340

[SYSTEM]: You are a helpful assistant. Your ultimate goal is to perform video temporal grounding (VTG) and referring video object segmentation (RefVOS). Based on the text query and video content, output your thinking process within `<think>` and `</think>` tags. If anything is unclear, you can select frames from the video for a clearer view by outputting `<select>{"start": start, "end": end, "keyframe": keyframe}</select>`, where 'start' and 'end' are the start and end frame indices of the region for detailed analysis, and 'keyframe' is the corresponding keyframe index. Once the final answer is confirmed, provide it within `<answer><VTG>{"start": start, "end": end, "keyframe": keyframe}</VTG>, <RefSeg><RefSeg></RefSeg></answer>`.

Sample 1 Expression: black SUV move in front of us

Round 1

[USER]: `<video>` Here are 10 low-resolution frames in this video (frame indice is [0, 1, 2, 4, 5, 6, 8, 9, 10, 12]), `<image>` Here are 5 high-resolution frames in this video (frame indice:[0, 3, 6, 9, 12]). The query is "black SUV move in front of us". Show me your thought process for finding and segmenting this target. This video has 13 frames.



Low-resolution Temporal Frames



Spatial Frames (with GT for clear visualization)

[Assistant]: `<think>` Step 1: The low-resolution overview shows a black SUV progressively moving ahead of the vehicle capturing this scene from frame 0 to 12. Step 2: Sampled frame 0 shows the black SUV behind other vehicles, frame 3 confirms its position, and frame 6 reveals it has advanced significantly. Step 3: The expression requires the black SUV to be the primary subject; the early frames show it overtaking others, while the final segment covers its continued forward motion. Action: I need to examine the 3-6 segment and keyframe 5 to verify the overtaking action. `</think><select><VTG>{"start": 3, "end": 6, "keyframe": 5}</VTG>, <RefSeg><SEG><RefSeg></select></im_end>`



Spatial Frames (Predictions)

Round 2

[USER]: `<video>` Here are 4 low-resolution frames in this video (frame indices are [3, 4, 5, 6]), `<image>` Here is key frames (frame indice: 5). The query is "black SUV move in front of us". Show me your thought process for finding and segmenting this target.



Selected Temporal Frames



Selected Key Frames

[Assistant]: `<think>` Step 1: The temporal segment 3-6 captures the black SUV actively overtaking the vehicle and becoming the leading car. Step 2: Keyframe 5 clearly shows the black SUV in front of the camera vehicle, confirming its position. Step 3: The overtaking action aligns with the expression's requirement for the black SUV to move in front. Action: The target is confirmed for the segment 0-12 with keyframe 6. `</think><answer><VTG>{"start": 0, "end": 12, "keyframe": 6}</VTG>, <RefSeg><SEG><RefSeg></answer></im_end>`



Keyframe GT

Keyframe Predict

Sample 2 Expression: A boy wearing a brown top, to the right of a boy wearing a black top and blue pants, on a stage.

Round 1

[USER]: `<video>` Here are 10 low-resolution frames in this video (frame indice is [0, 30, 61, 91, 122, 152, 183, 213, 244, 275]), `<image>` Here are 5 high-resolution frames in this video (frame indice:[0, 68, 137, 206, 275]). The query is "A boy wearing a brown top, to the right of a boy wearing a black top and blue pants, on a stage.". Show me your thought process for finding and segmenting this target. This video has 276 frames.



Low-resolution Temporal Frames



Spatial Frames (with GT for clear visualization)

[Assistant]: `<think>` Step 1: The low-resolution overview shows a room with two girls interacting; one is frequently positioned in front of the other. Step 2: In sampled frames 0 and 65, the girl with short hair is seen standing while the other girl crouches; in later frames (131+), the roles reverse. Step 3: The expression specifies a short-haired girl with another girl beside her, which aligns with the persistent pair interaction observed throughout the video. Action: I need to examine the 16-108 segment and keyframe 66 to verify the initial interaction pattern. `</think><select><VTG>{"start": 16, "end": 108, "keyframe": 66}</VTG>, <RefSeg><SEG><RefSeg></select></im_end>`



Spatial Frames (Predictions)

Round 2

[USER]: `<video>` Here are 5 low-resolution frames in this video (frame indices are [16, 39, 62, 85, 108]), `<image>` Here is key frames (frame indice: 66). The query is "A boy wearing a brown top, to the right of a boy wearing a black top and blue pants, on a stage.". Show me your thought process for finding and segmenting this target.

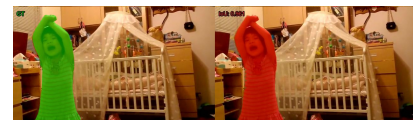


Selected Temporal Frames



Selected Key Frames

[Assistant]: `<think>` Step 1: The temporal segment 16-108 captures the girls' dynamic interaction, including the short-haired girl crouching while the other stands. Step 2: Keyframe 66 shows the short-haired girl in a crouched position with the other girl positioned directly beside her. Step 3: The visual evidence confirms the short-haired girl with the other girl beside her, matching the expression's description. Action: The target is confirmed. `</think><answer><VTG>{"start": 16, "end": 108, "keyframe": 66}</VTG>, <RefSeg><SEG><RefSeg></answer></im_end>`



Keyframe GT

Keyframe Prediction

Figure 9. Visualization of a two-round interactive reasoning and segmentation process.

1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484

[SYSTEM]: You are a helpful assistant. Your ultimate goal is to perform video temporal grounding (VTG) and referring video object segmentation (RefVOS). Based on the text query and video content, output your thinking process within <think> and </think> tags. If anything is unclear, you can select frames from the video for a clearer view by outputting <select>{"start": start, "end": end, "keyframe": keyframe}</select>, where 'start' and 'end' are the start and end frame indices of the region for detailed analysis, and 'keyframe' is the corresponding keyframe index. Once the final answer is confirmed, provide it within <answer><VTG>{"start": start, "end": end, "keyframe": keyframe}</VTG>, <RefSeg></RefSeg></answer>.

Sample 1 Expression: a stopped car start moving forward fast and bypass another white car

Round 1

[USER]: <video> Here are 10 low-resolution frames in this video (frame indices are [0, 4, 8, 13, 17, 21, 26, 30, 34, 39]). <image> Here are high-resolution frames (frame indices: [0, 9, 19, 29, 39]). The query is "a stopped car start moving forward fast and bypass another white car". Show me your thought process for finding and segmenting this target.



Low-resolution Temporal Frames



Spatial Frames (with GT for clear visualization)

[Assistant]: <think> Step 1: The low-resolution overview shows the video beginning with a stationary black SUV followed by other vehicles, then transitioning to an intersection where traffic flow changes. Step 2: Sampled frame 0 shows the SUV stopped; frame 9 shows it has moved forward and is approaching another car; frames 19-39 depict continued movement with vehicles shifting positions. Step 3: The expression requires identifying a car that starts moving forward quickly and overtakes a white car, which likely occurs early in the video. Action: I need to examine the 0-12 segment and keyframe 4 to verify the initial movement. </think><select><VTG>{"start": 0, "end": 12, "keyframe": 4}</VTG>, <RefSeg>[SEG]</RefSeg></select></im_end>



Spatial Frames (Predictions)

Round 2

[USER]: <video> Here are 5 low-resolution frames in this video (frame indices are [0, 3, 6, 9, 12]). <image> Here is key frames (frame indice: 4). The query is "a stopped car start moving forward fast and bypass another white car". Show me your thought process for finding and segmenting this target.



Selected Temporal Frames



Selected Key Frames

[Assistant]: <think> Step 1: The temporal segment 0-12 captures the black SUV initiating movement, but the white car it overtakes becomes visible only in later frames. Step 2: Keyframe 4 confirms the black SUV remains stationary while a white car in front remains stationary, matching the 'stopped car start moving' condition. Step 3: The critical 'bypassing a white car' action occurs later (around frame 9-12), indicating the expression's target is not fully contained in this segment. Action: I will now examine the 8-18 segment and keyframe 10 to verify the overtake event. </think><select><VTG>{"start": 8, "end": 18, "keyframe": 10}</VTG>, <RefSeg>[SEG]</RefSeg></select></im_end>



Keyframe GT

Keyframe Predict

Round 3

[USER]: <video> Here are 5 low-resolution frames in this video (frame indices are [8, 10, 13, 15, 18]). <image> Here is key frames (frame indice: 10). The query is "a stopped car start moving forward fast and bypass another white car". Show me your thought process for finding and segmenting this target.



Selected Temporal Frames



Selected Key Frames

[Assistant]: <think> Step 1: The temporal segment 8-18 shows the black SUV gradually accelerating and passing the white car in front. Step 2: Keyframe 10 clearly depicts the black SUV starting to overtake the white car as they move forward together. Step 3: The expression's core action of the black car moving quickly to overtake the white car is fully contained within this segment, with the initial acceleration visible by frame 8. Action: Based on this complete sequence, the target is confirmed. </think><answer><VTG>{"start": 8, "end": 18, "keyframe": 10}</VTG>, <RefSeg>[SEG]</RefSeg></answer></im_end>



Keyframe GT

Keyframe Predict

Figure 10. Visualization of a **three-round** iterative refinement process for complex target localization (Sample 1).

[SYSTEM]: You are a helpful assistant. Your ultimate goal is to perform video temporal grounding (VTG) and referring video object segmentation (RefVOS). Based on the text query and video content, output your thinking process within <think> and </think> tags. If anything is unclear, you can select frames from the video for a clearer view by outputting <select>{"start": start, "end": end, "keyframe": keyframe}</select>, where 'start' and 'end' are the start and end frame indices of the region for detailed analysis, and 'keyframe' is the corresponding keyframe index. Once the final answer is confirmed, provide it within <answer><VTG>{"start": start, "end": end, "keyframe": keyframe}</VTG>, <RefSeg></RefSeg></answer>.

Sample 2 Expression: cow standing then walking to the left

Round 1

[USER]: <video> Here are 10 low-resolution frames in this video (frame indices are [0, 13, 26, 39, 52, 66, 79, 92, 105, 119]), <image> Here are high-resolution frames (frame indices: [0, 29, 59, 89, 119]). The query is "cow standing then walking to the left". Show me your thought process for finding and segmenting this target.



Low-resolution Temporal Frames



Spatial Frames (with GT for clear visualization)

[Assistant]: <think> Step 1: The low-resolution overview shows multiple cows in a grassy field; some stand stationary while others move leftward. Step 2: Sampled frame 0 shows a cow standing on the right, frame 29 shows a cow walking left, indicating a leftward movement phase. Step 3: The expression requires a cow that stands then walks left; the first segment (0-40) covers the transition from standing to walking left. Action: I need to examine the 0-40 segment and keyframe 20 to verify the standing-to-walking transition. <think><select><VTG>{"start": 0, "end": 40, "keyframe": 20}</VTG>, <RefSeg>[SEG]</RefSeg></select></im_end>



Spatial Frames (Predictions)

Round 2

[USER]: <video> Here are 5 low-resolution frames in this video (frame indices are [0, 10, 20, 30, 40]), <image> Here is key frames (frame indice: 10). The query is "cow standing then walking to the left". Show me your thought process for finding and segmenting this target.



Selected Temporal Frames



Selected Key Frames

[Assistant]: <think> Step 1: The 0-40 segment captures the cow moving left from a stationary position. Step 2: Keyframe 20 shows the cow standing and turning left, with the head visible in the grass. Step 3: The expression is satisfied by this segment, but the Final Segment (15-60) better aligns with the walking phase as the cow remains positioned on the left side. Action: The target is confirmed within the Final Segment (15-60). <think><select><VTG>{"start": 15, "end": 60, "keyframe": 35}</VTG>, <RefSeg>[SEG]</RefSeg></select></im_end>



Keyframe GT

Keyframe Predict

Round 3

[USER]: <video> Here are 5 low-resolution frames in this video (frame indices are [15, 26, 37, 48, 60]), <image> Here is key frames (frame indice: 35). The query is "a stopped car, start moving forward fast and bypass another white car". Show me your thought process for finding and segmenting this target.



Selected Temporal Frames



Selected Key Frames

[Assistant]: <think> Step 1: The Final Segment (15-60) shows the cow moving left while remaining stationary on the left side. Step 2: Keyframe 35 confirms the cow is walking left. Step 3: The Final Segment better captures the walking phase, but the initial standing phase starts at frame 0. The first segment (0-40) is the most accurate for the full expression. Action: The target is confirmed with the Final Segment. <think><select><VTG>{"start": 0, "end": 40, "keyframe": 20}</VTG>, <RefSeg>[SEG]</RefSeg></select></im_end>



Keyframe GT

Keyframe Predict

Figure 11. Visualization of a **three-round** iterative refinement process for complex target localization (Sample 2).