VideoSeg-R1:Reasoning Video Object Segmentation via Reinforcement Learning

Zishan Xu[1,2,*], **Yifu Guo**[1,3,*], **Yuquan Lu**[1,3,*], **Fengyu Yang**[1],
Junxin Li[1,3], **Lihua Cai**[1,4,†]

[1]Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China
[2]School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[3]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[4]Xiamen Rekey Medical Technology Co., LTD
xuzishan@m.scnu.edu.cn, lee.cai@m.scnu.edu.cn
[*]Equal contribution. [†]Corresponding author.

Abstract

Traditional video reasoning segmentation methods rely on supervised fine-tuning, which limits generalization to out-of-distribution scenarios and lacks explicit reasoning. To address this, we propose **VideoSeg-R1**, the first framework to introduce reinforcement learning into video reasoning segmentation. It adopts a decoupled architecture that formulates the task as joint referring image segmentation and video mask propagation. It comprises three stages: (1) A hierarchical text-guided frame sampler to emulate human attention; (2) A reasoning model that produces spatial cues along with explicit reasoning chains; and (3) A segmentation-propagation stage using SAM2 and XMem. A task difficulty-aware mechanism adaptively controls reasoning length for better efficiency and accuracy. Extensive evaluations on multiple benchmarks demonstrate that VideoSeg-R1 achieves state-of-the-art performance in complex video reasoning and segmentation tasks.

Code — https://github.com/euyis1019/VideoSeg-R1

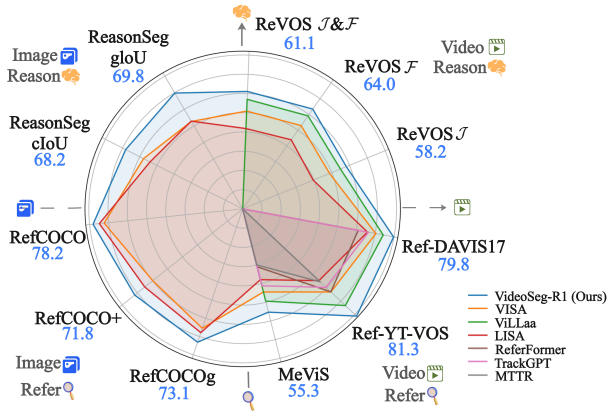

Figure 1. VideoSeg-R1 achieves state-of-the-art performance on both video and image benchmarks covering reasoning and referring segmentation tasks.

Introduction

Referring video object segmentation (RVOS) requires a model to *localize* and *segment* **one or multiple target objects** throughout an entire video, given a natural-language description (Guo, Wang, and Zhang 2019). Success hinges on two intertwined capabilities: (i) fine-grained *spatial* precision at the pixel level and (ii) robust *temporal* reasoning to track objects under motion, occlusion, and appearance change. In recent years, large language models (LLMs) have made remarkable progress across various dimensions (Luo et al. 2025; Lin et al. 2025; Du, Liu, and Zhang 2025b,a; Du et al. 2025), which in turn has driven the advancement of multimodal models. While recent multimodal large language models (MLLMs) excel on static-image tasks, they falter when confronted with long-form videos and complex language queries that demand multi-step reasoning.

Existing RVOS pipelines overwhelmingly rely on **supervised fine-tuning (SFT)** (Wu et al. 2022a). Although effective on in-distribution data, SFT models (i) overfit to seen categories and viewpoints (Zhang et al. 2025), (ii) lack interpretable reasoning chains. Consequently, accuracy plummets when a query involves subtle temporal context (e.g., "*the man who appears after the car turns left*") or commonsense inference (Bellver et al. 2022).

Reinforcement learning (RL) has recently emerged as a powerful tool for endowing language models with reasoning abilities. Algorithms such as Group Relative Policy Optimization (GRPO) (Shao et al. 2024) have advanced chain-of-thought generation, and their multimodal extensions have begun to tackle pixel-level vision tasks. *Yet no prior work* has transferred RL-based reasoning to the video reasoning segmentation domain, where the action space explodes with temporal length and the reward must balance spatial accuracy against temporal consistency.

We bridge this gap with **VideoSeg-R1**—the first framework that casts RVOS as an *RL-driven, reasoning-centric* problem, capable of handling multi-target queries. VideoSeg-R1 adopts a *decoupled* three-stage design:

1. *Hierarchical text-guided frame sampling* progressively narrows the search space, mimicking human coarse-to-fine attention to effectively isolate key clips while reducing redundancy.

[*]Equal contribution
[†]Corresponding author

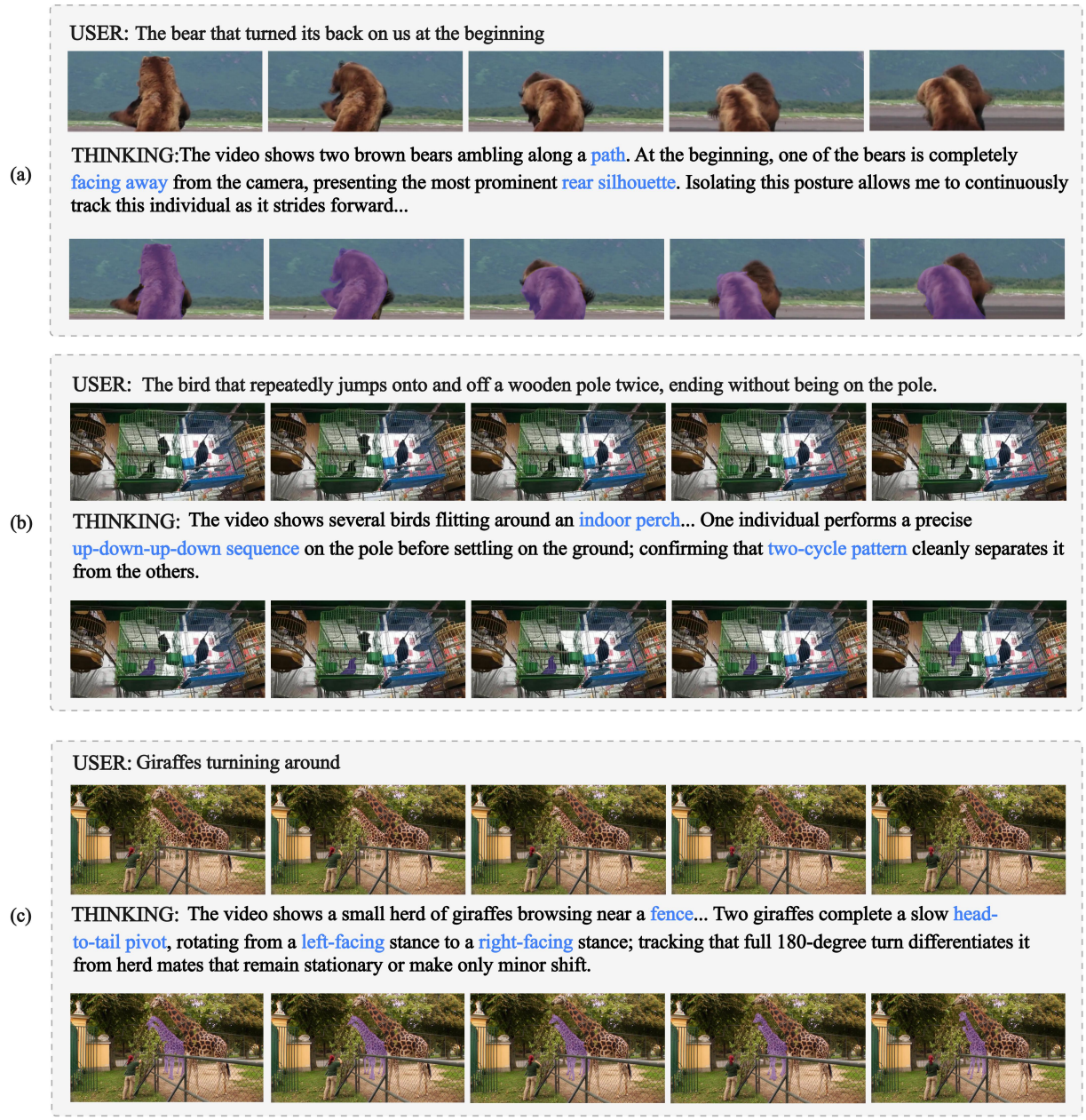

Figure 2. Our VideoSeg-R1 effectively segments and tracks in challenging scenarios, including: (a) objects in crowded scenes; (b) multiple objects with rapid motion; and (c) diverse targets appearing simultaneously.

2. *GRPO-enhanced multimodal reasoning* operates on the selected frames, generating explicit spatial cues (bboxes & points) together with a chain of thought whose length is modulated by a *task-difficulty-aware soft penalty*.
3. *Seg-prop decoupling* sends the sparse cues to state-of-the-art segmentation (SAM2) and bidirectional propagation (XMem) modules, producing pixel-accurate masks for every frame at a fraction of the computation cost.

Extensive experiments on Ref-YouTube-VOS (Seo, Lee, and Han 2020), MeViS (Ding et al. 2023), Ref-DAVIS17 (Pont-Tuset et al. 2018) and ReVOS (Yan et al. 2024) show that VideoSeg-R1 sets a new state of the art, with particularly large gains ($\geq$ 6.0% J&F) on reasoning-intensive queries. Ablations confirm that 1.hierarchical sampling improves performance by precisely locating key frames, 2.difficulty-aware GRPO boosts reasoning fidelity, and 3.decoupling reasoning from propagation is essential for temporal stability.

Our Contributions are listed below:

- We introduce **the first RL framework for reasoning-aware RVOS**, unifying explicit chain-of-thought generation with temporal mask propagation.

- We devise a **hierarchical frame sampler** that aligns computational effort with query semantics, enabling efficient processing of minute-long videos.
- We propose a **task-difficulty-aware soft length penalty** that adaptively controls reasoning depth under GRPO, improving both accuracy and efficiency.
- We achieve new **state-of-the-art** results on multiple benchmarks, validating the effectiveness and generality of VideoSeg-R1.

By marrying reinforcement learning with video segmentation, VideoSeg-R1 opens a new research avenue for *explicitly interpretable, resource-aware* video understanding.

Related Work

Multi-Modal Large Language Model

Multimodal Large Language Models (MLLMs) have significantly advanced vision-language tasks in recent years, with notable examples including InstructBLIP (Dai et al. 2023), InternGPT (Liu et al. 2023), QwenVL (Bai et al. 2025), and Intern-Video2 (Wang et al. 2024). Despite their success, existing MLLMs still have considerable scope for improvement in reasoning abilities. To enhance these capabilities, methods such as process-based reward models (Lightman et al. 2023; Uesato et al. 2022), reinforcement learning (Kumar et al. 2024) and search algorithms (Feng et al. 2023; Trinh et al. 2024) have been explored. Among these, DeepSeek R1, trained with the GRPO algorithm, has demonstrated strong reasoning performance and test-time scalability. Building on this, reinforcement learning techniques have recently been applied to multimodal large language models. Examples include Open-R1-Multimodal (Lab 2025), emphasizing mathematical reasoning, and R1-V (R1-V Team 2025), excelling at counting tasks. Additionally, studies such as Seg-Zero, SAM-R1, and Seg-R1 have targeted fine-grained pixel-level understanding. However, current research primarily addresses static image scenarios and lacks comprehensive temporal reasoning.

Extending these models to video domains introduces significant challenges in managing temporal dimensions. Issues like long-term video perception, language ambiguity, object occlusion, rapid motion, and appearance variations complicate temporal reasoning. To bridge this gap, we propose VideoSeg-R1, which integrates reinforcement learning into the Reasoning Video Object Segmentation task for the first time, significantly enhancing pixel-level temporal reasoning capabilities in video scenarios.

MLLMs for Segmentation

Early methods like LISA (Lai et al. 2024) introduced a special `<SEG>` token, bridging MLLMs with segmentation models such as SAM to produce accurate segmentation masks. Following this paradigm, methods like PixelLM (Ren et al. 2024), GLaMM (Rasheed et al. 2024), and TEXT4SEG (Lan et al. 2024) focused primarily on static image segmentation, exhibiting limited adaptability for video object segmentation.

For example, TrackGPT (Stroh 2024b) extended LISA by updating tokens iteratively across video frames, yet ignored temporal dependencies. VISA addressed this with keyframe sampling to handle multiple frames but suffered from cumulative errors due to inaccurate keyframe selection. VideoLISA reduced computational load through sparse sampling but lacked adaptive keyframe extraction, causing redundancy. Although ViLLa improved sampling accuracy with key segment extraction, it faced significant computational overhead in long videos or lengthy action sequences. Moreover, these methods typically rely on SFT, limiting generalization to out-of-distribution (OOD) samples and causing catastrophic forgetting, hindering real-world applicability.

To address these issues, we propose VisionSeg-R1, featuring a decoupled design (Guo et al. 2025; Lai et al. 2024) and a Hierarchical Text-guided Frame Sampler that mimics human attention strategies to effectively reduce redundancy. Additionally, we leverage the GRPO algorithm to enhance reasoning capabilities and generalization performance.

Method

Pipeline Formulation

Given a text query instruction x_t and an input video $x_v = \{f_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$ with T frames, our goal is to design a model $\phi_\theta(\cdot)$ that generates a binary segmentation mask sequence $M = \{m_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W}$, which precisely localizes the target object in the video based on the query semantics.

$$M = \phi_\theta(x_t, x_v) \tag{1}$$

The complexity of the text query x_t varies: it may be a simple phrase that directly describes the appearance, action, or position of the target (e.g., "the woman in red"), or it could involve expressions that require world knowledge or commonsense reasoning (e.g., "the person who looks like a doctor"). It may also require complex inference about the video content and future developments (e.g., "who is most likely to be the main character in this wedding?"). The latter two types of queries demand more advanced semantic understanding and reasoning capabilities from the model.

To address these challenges and leverage recent advances in the reasoning abilities of LLM, we propose a reasoning-based video object segmentation pipeline. We formulate the overall task as a joint problem of referring image segmentation and video mask propagation, consisting of the following three stages: (1) A hierarchical text-guided frame sampler to emulate human attention; (2) A reasoning model that produces spatial cues along with explicit reasoning chains; and (3) A segmentation-propagation stage using SAM2 and XMem.

VideoSeg-R1 Model

Hierarchical Text-guided Frame Sampler. To better align with human perception in long video understanding, we propose a hierarchical text-guided sampling strategy that simulates progressive attention convergence. In long video understanding, humans typically begin with a coarse estimate of when an event may occur, and progressively refine their attention to identify the precise frame. Inspired by this, we design a multi-round reasoning structure that iteratively compresses semantics to locate key segment and traget frame.

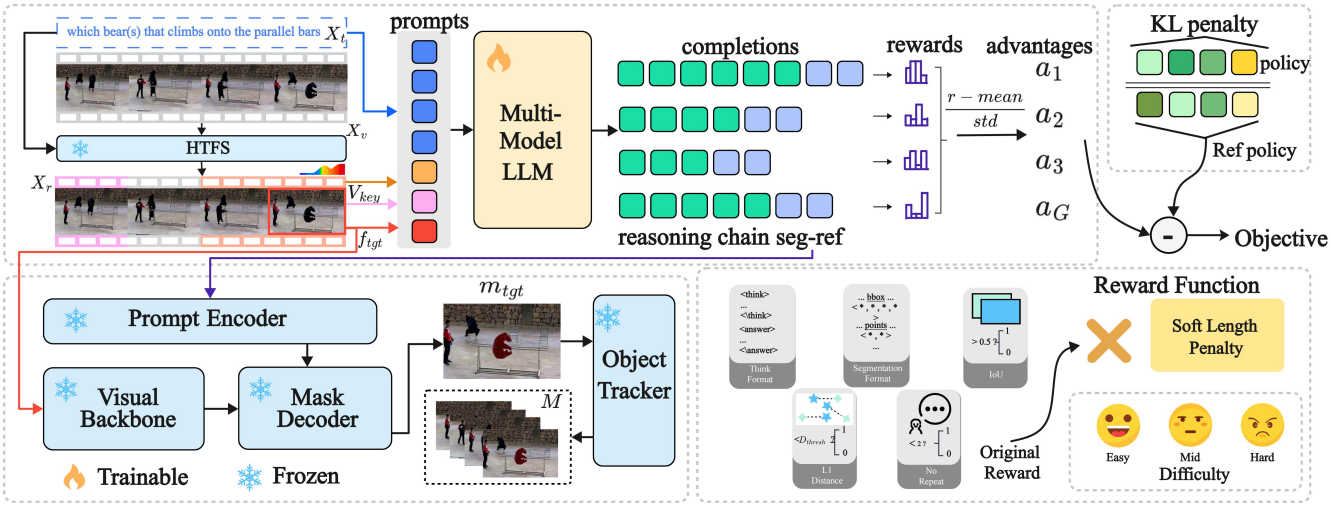

Figure 3. Overview of VideoSeg-R1, which consists of the following three stages: (1) a hierarchical text-guided frame sampler to emulate human attention; (2) a reasoning model that produces spatial cues along with explicit reasoning chains; and (3) a segmentation-propagation stage using SAM2 and XMem.

Specifically, given an input video $x_v \in \mathbb{R}^{T\times H\times W\times 3}$ and a textual query x_t, we first prompt a video understanding LLM (e.g., Qwen2.5-VL (Bai et al. 2025)) to predict the time intervals semantically relevant to the query. We extract K temporal intervals $(t_i^{\text{start}}, t_i^{\text{end}})$ from multiple responses and compute their average to determine the key segment boundaries $[t_{\text{key}}^{\text{start}},\ t_{\text{key}}^{\text{end}}]$. The corresponding set of frames is denoted as $V_{\text{key}} = \{f_{t_{\text{key}}^{\text{start}}}, f_{t_{\text{key}}^{\text{start}}+1}, \ldots, f_{t_{\text{key}}^{\text{end}}}\}$, has a length of T_{key}. To prevent errors caused by performing frame-level localization on overly long segments, we introduce a relative segment length threshold $\delta \in (0, 1)$. When the key segment length $T_{\text{key}} > \delta \cdot T$, we continue applying semantic compression to V_{key} until $T_{\text{key}} \le \delta \cdot T$, indicating that the model's attention has sufficiently converged for frame-level reasoning. In the frame-level localization phase, the model predict the approximate percentage position of the target frame within the key segment.

We collect the top-M predicted values $\{p_i\}_{i=1}^{M}$ and compute the average $\bar{p}$ to determine the final target frame $f_{\text{tgt}} = f_{t_{\text{key}}^{\text{start}}+\lfloor T_{\text{key}}\cdot\bar{p}\rfloor}$. To implement the semantic convergence process described above, we introduce two stage-specific prompt templates. See the Section A for details. To provide global context, we also employ an adaptive global sampling strategy to sample reference frames X_r. Different sampling strategies are detailed in the Section B.

Reasoning Model. We adopt Qwen2.5-VL (Bai et al. 2025) as the reasoning model F_{reason}. At this stage, the model takes as input the key segment frames V_{key}, the target frame f_{tgt}, and the reference frame set X_r, along with the textual query x_t, and performs multimodal reasoning.

During reinforcement learning, the model is optimized to generate structured outputs. These outputs are parsed by a post-processing function G to extract the target bounding box B, a central point P_{central}, and a negative point P_{neg}, which help improve spatial precision and distinguish between multiple objects. Each point is represented as $P = (x, y, l)$, where (x, y) denotes the spatial coordinates, and $l \in \{0, 1\}$ is a binary label indicating whether the point is positive ($l = 1$) or negative ($l = 0$). This process can be formally expressed as:

$$B,\ P_{\text{central}},\ P_{\text{neg}} = G(F_{\text{reason}}(V_{\text{key}},\ f_{\text{tgt}},\ X_r,\ x_t)) \quad (2)$$

Segmentation and propagation stage. We employ SAM2 (Ravi et al. 2024) as the segmentation model F_{seg}, chosen for its high accuracy and efficient inference. Given the target frame f_{tgt}, a visual backbone E_v first extracts its features. The predicted bounding box B, together with a positive point P_{central} and a negative point P_{neg}, serves as spatial prompts to guide the segmentation model in generating the target mask:

$$m_{\text{tgt}} = \text{SAM2}(E_v(f_{\text{tgt}}), B, P_{\text{central}}, P_{\text{neg}}) \quad (3)$$

To extend the segmentation across the entire video, we apply XMem (Cheng and Schwing 2022), an advanced object tracking model, to propagate m_{tgt} bidirectionally:

$$M = \{m_t\}_{t=1}^{T} = \text{OT}(m_{\text{tgt}}, x_v) \quad (4)$$

Reward Functions

Original Reward. We adopt the original reward design in VisionReasoner (Liu et al. 2025). The overall reward function consists of the following components:

$$R_{\text{original}} = R_{\text{format}} + R_{\text{seg\_accuracy}} + R_{\text{non\_repeat}} \quad (5)$$

$$R_{\text{format}} = R_{\text{reason\_format}} + R_{\text{seg\_format}} \quad (6)$$

which assess the reasoning format, segmentation format, segmentation accuracy, and non-redundant reasoning, respectively. $R_{\text{non\_repeat}}$ measures the diversity of the reasoning process by assigning higher rewards to outputs composed of unique, non-repetitive sentences. The segmentation accuracy reward $R_{\text{seg\_accuracy}}$ comprises the evaluation of mask

IoU, point-level L1 distance, and bounding box-level L1 distance. See Appendix C for details.

Negative Point Distance Reward. To enhance the model's ability to distinguish foreground from background, we introduce a negative point distance reward. For each predicted negative point $(x, y, 0)$, we compute its minimum L1 distance to all ground-truth target regions (annotated masks). If the distance is positive and does not exceed the predefined threshold τ_{neg} (40 pixels), the point is considered valid and receives a reward increment of $\frac{1}{K}$, where K is the number of predicted negative points; if the distance is zero or negative—that is, if the point lies inside or on the boundary of the target mask—no reward is given. This mechanism encourages the model to place negative points close to, but outside, the target regions, thereby improving foreground–background separability.

Task Difficulty. To enable efficient training for video reasoning segmentation, we estimate an instance-level task difficulty score $D \in [1, 10]$ for each sample. Specifically, we prompt a MLLM to rate the sample across five dimensions: *scene complexity*, *segmentation challenge*, *temporal ambiguity*, *motion complexity*, and *linguistic ambiguity*. Each dimension is scored on a 1–10 scale, and the final difficulty score is computed as the average of these five ratings. We also categorize difficulty into three levels (easy, medium, and hard) using thresholds τ_{easy} and τ_{hard}. This difficulty prior is then used to guide reasoning token budget allocation in reinforcement learning, allowing the model to adaptively adjust reasoning length based on the sample's difficulty. Details are provided in the Section D.

Soft Length Penalty. To enable adaptive control over reasoning length under varying task complexities, we propose a soft length penalty mechanism. Unlike traditional methods that rely on hard truncation, our approach encourages concise reasoning for simple tasks while allowing more detailed reasoning for complex ones. Specifically, we define the expected reasoning token budget L_{budget} based on the task difficulty score D as follows: $L_{\text{budget}} = L(D)$, where $L(D)$ denotes the base budget allocated according to task difficulty D; the detailed mapping strategy is provided in Section E. Let L_{used} be the number of reasoning tokens actually generated by the model. When it exceeds the budget, a soft penalty is applied to the reward:

$$s(L_{\text{used}}, L_{\text{budget}}) = \begin{cases} 1 - \beta \cdot (L_{\text{used}} - L_{\text{budget}}), \\ \qquad\qquad \text{if } L_{\text{used}} > L_{\text{budget}} \\ 1, \quad \text{otherwise} \end{cases} \tag{7}$$

The final reward is computed as:

$$R_{\text{final}} = R_{\text{original}} \cdot s(L_{\text{used}}, L_{\text{budget}}) \tag{8}$$

Multi-object Matching

To support multi-object segmentation, our framework employs batched computation and the Hungarian algorithm to effectively handle the many-to-many matching problem under bounding box IoU reward, bounding box L1 reward, and center point L1 reward. For each observed object o_j, we maintain a list of predicted bounding boxes $(B^i_{\text{pred}})_{i=1}^K$ and predicted center points $(P^i_{\text{pred}})_{i=1}^K$, and compute the reward scores in batch with respect to the ground-truth bounding boxes $(B^i_{\text{GT}})_{i=1}^N$ and ground-truth center points $(P^i_{\text{GT}})_{i=1}^N$. Subsequently, we use the Hungarian algorithm to compute the optimal one-to-one assignment. This design ensures both optimal alignment between predictions and ground-truth annotations and efficient computation performance.

Experiment

Dataset

Training Dataset. We train VisionSeg-R1 using the Ref-YouTube-VOS (Seo, Lee, and Han 2020), MeViS (Ding et al. 2023) and Ref-DAVIS17 (Pont-Tuset et al. 2018). For the mask annotations in Referring VOS datasets, we extract the leftmost, topmost, rightmost, and bottommost pixels of each target mask to generate the bounding box B. In addition, we compute the center point coordinates of each mask. To support multi-object expressions, we handle multiple objects per image by: (i) using one center point per object (ii) assembling all corresponding bounding boxes and center points into respective lists.

Evaluation Dataset. For evaluation, we test the model on both video and image datasets: (1) For video datasets, we use the ReVOS (Yan et al. 2024) dataset to evaluate the performance of ReasonVOS, and the Ref-YouTube-VOS, MeViS and Ref-DAVIS17 to evaluate the performance of vanilla Referring VOS performance. (2) For image datasets, we use ReasonSeg (Lai et al. 2024), refCOCO (Kazemzadeh et al. 2014), refCOCO+ (Kazemzadeh et al. 2014), and refCOCOg (Mao et al. 2016) to evaluate the generalization ability of the VisionSeg-R1 model on image-level segmentation tasks.

Implementation Details

We adopt Qwen2.5-VL-7B and Qwen2.5-VL-3B (Bai et al. 2025) as our reasoning models and video understanding models, and use SAM2-Large (Ravi et al. 2024) as the default segmentation model. In addition, we utilize XMem (Cheng and Schwing 2022), a semi-supervised video object segmentation method, as the object tracker. We train VisionSeg-R1 using the DeepSpeed (Rasley et al. 2020) library on 8 NVIDIA 80G A100 GPUs. During training, the Hierarchical Text-guided Frame Sampler, the visual backbone, the SAM2 decoder, the prompt encoder, and the object tracker are all frozen. Only the multi-modal LLM is updated. Note that the Hierarchical Text-guided Frame Sampler is only used during inference. During training, we use a total batch size of 16 with a sampling number of 8 per training step. The initial learning rate is set to 1e-6 and the weight decay is 0.01. More training details are provided in the Section F.

Evaluation Metrics

For image-based evaluation, we adopt two commonly used metrics: gIoU and cIoU, following prior works (Kazemzadeh et al. 2014; Lai et al. 2024). Specifically, gIoU

Method	Ref-YouTube-VOS			Ref-DAVIS17			MeViS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS (Seo, Lee, and Han 2020)	47.2	45.3	49.2	51.6	47.3	56.0	27.8	25.7	29.9
MTTR (Botach, Zheltonozhskii, and Baskin 2022)	55.3	54.0	56.6	-	-	-	30.0	28.8	31.2
LBDT (Ding et al. 2022)	49.4	48.2	50.6	54.1	-	-	29.3	27.8	30.8
ReferFormer (Wu et al. 2022b)	62.9	61.3	64.1	61.1	58.1	64.1	31.0	29.8	32.2
LMPM (Ding et al. 2023)	-	-	-	-	-	-	37.2	34.2	40.2
OnlineRefer (Wu et al. 2023)	62.9	61.0	64.7	62.4	59.1	65.6	-	-	-
DsHmp (He and Ding 2024)	67.1	65.0	69.1	64.9	61.7	68.1	46.4	43.0	49.8
TrackGPT (Stroh 2024a)	59.5	58.1	60.8	66.5	62.7	70.4	41.2	39.2	43.1
LISA (Lai et al. 2024)	54.4	54.0	54.8	66.0	63.2	68.8	37.9	35.8	40.0
PixelLM (Ren et al. 2024)	55.0	54.3	55.7	66.7	63.4	70.0	38.7	36.3	41.1
VideoLISA (Fu et al. 2025)	61.7	60.2	63.1	67.7	63.8	71.5	42.3	39.4	45.2
VISA (Yan et al. 2024)	63.0	61.4	64.6	70.4	66.7	73.8	44.5	41.8	47.1
ViLLa (Zheng et al. 2025)	73.3	70.5	76.6	74.3	70.6	78.0	49.4	46.5	52.3
VideoSeg-R1(Qwen2.5-VL-3B)	**75.9**	**73.3**	**78.5**	**77.5**	**74.4**	**80.5**	**53.0**	**50.7**	**55.3**
VideoSeg-R1(Qwen2.5-VL-7B)	**81.3**	**78.2**	**84.4**	**79.8**	**77.4**	**82.2**	**55.3**	**52.7**	**57.8**

Table 1. Referring Video Object Segmentation on Ref-YouTube-VOS, Ref-DAVIS17, and MeViS.

Methods	Backbone	refCOCO			refCOCO+			refCOCOg		ReasonSeg	
		val	testA	testB	val	testA	testB	val(U)	test(U)	gIoU	cIoU
MCN (Luo et al. 2020)	Darknet53	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4	-	-
VLT (Ding et al. 2021)	Darknet53	65.7	68.3	62.7	55.5	59.2	49.4	53.0	56.7	-	-
CRIS (Wang et al. 2022)	ResNet101	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	-	-
LAVT (Yang et al. 2022)	Swin-B	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	-	-
ReLA (Liu, Ding, and Jiang 2023)	Swin-B	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	-	-
X-Decoder (Zou et al. 2023a)	DaViT-L	-	-	-	-	-	-	64.6	-	22.6	17.9
SEEM (Zou et al. 2023b)	DaViT-L	-	-	-	-	-	-	65.7	-	25.5	21.2
LISA (Lai et al. 2024)	LLaVA-7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	52.9	54.0
VISA (Yan et al. 2024)	Chat-UniVi-7B	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	52.7	57.8
VideoLISA (Fu et al. 2025)	LLaVA-Phi-3-V-3.8B	73.8	76.6	68.8	63.4	68.8	56.2	68.3	68.8	61.4	67.1
VideoSeg-R1 (Ours)	**Qwen2.5-VL-3B**	**75.1**	**79.2**	**72.8**	**67.2**	**72.8**	**59.9**	**69.7**	**71.0**	**65.1**	**63.7**
VideoSeg-R1 (Ours)	**Qwen2.5-VL-7B**	**78.2**	**82.3**	**75.1**	**71.8**	**76.1**	**64.7**	**73.1**	**74.1**	**69.8**	**68.2**

Table 2. Referring image segmentation on the refCOCO, refCOCO+, refCOCOg, and ReasonSeg datasets.

Method	ReVOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
LISA-LLAVA-13B	41.8	39.6	43.9
VISA-Chat-UniVi-13B	50.9	48.8	52.9
VISA-InternVideo2-6B	52.4	50.1	54.7
ViLLa-InternVideo2-6B	57.0	54.9	59.1
VideoSeg-R1-Qwen2.5-VL-3B	**58.6**	**56.4**	**60.8**
VideoSeg-R1-Qwen2.5-VL-7B	**61.1**	**58.2**	**64.0**

Table 3. Video Reasoning Segmentation on ReVOS.

is computed as the average of all per-image Intersection-over-Unions (IoUs), while cIoU is defined by the cumulative intersection over the cumulative union. For video-based evaluation, we follow previous studies (Wu et al. 2022b, 2023), using region similarity (J), contour accuracy (F), and their average value (J&F).

Model	Type	CoT	MeViS J&F	ReVOS J&F	ReasonSeg gIoU
Baseline			38.1	40.2	56.1
VideoSeg-R1	SFT	✗	45.4	50.9	59.4
VideoSeg-R1	RL	✗	52.6	58.9	67.2
VideoSeg-R1	RL	✓	**55.3**	**61.1**	**69.8**

Table 4. Performance comparison between SFT and RL.

Comparison

We compare our model with prior works through quantitative evaluations on standard benchmarks (Tables 1, 2, 3) and qualitative comparisons provided in the Section G.

Referring VOS. In the task of referring video object segmentation, VideoSeg-R1 achieves leading performance on three standard benchmarks: Ref-YouTube-VOS, Ref-DAVIS17, and MeViS, with J&F scores of 81.3, 79.8, and 55.3 respectively. These results significantly surpass mainstream methods such as ViLLa (73.3, 74.3, 49.4) and VISA (63.0, 70.4, 44.5), as shown in Table 1.

ID	Key Seg.	Target Fr.	MeViS	ReVOS
1	✗	✗	45.7	53.9
2	✗	✓	52.5	57.6
3	✓	✗	49.2	56.3
4	✓	✓	**55.3**	**61.1**

Table 5. Ablation study of target localization strategies.

Model	Bbox	P_{central}	P_{neg}	ReVOS J&F	ReasonSeg gIoU
Baseline				40.2	56.1
VideoSeg-R1	✗	✓	✗	56.5	64.0
VideoSeg-R1	✓	✗	✗	57.2	65.7
VideoSeg-R1	✓	✓	✗	60.3	68.2
VideoSeg-R1	✓	✗	✓	58.1	66.5
VideoSeg-R1	✓	✓	✓	**61.1**	**69.8**

Table 6. Ablation study on the effect of Bbox, P_{central}, and P_{neg} as spatial prompts.

Image Datasets. Images can be treated as single-frame videos, allowing VideoSeg-R1 to be directly applied to image datasets without any modification. As shown in Table 2, our method consistently outperforms existing state-of-the-art approaches across all benchmarks. On refCOCO, VideoSeg-R1 achieves a validation score of 78.2, surpassing LISA and VideoLISA by 3.3 and 4.4 points, respectively. For refCOCO+ and refCOCOg, our model obtains scores of 71.8 and 73.1, outperforming LISA by 6.7 and 5.2 points, respectively. Most notably, on the reasoning-intensive ReasonSeg dataset, VideoSeg-R1 achieves 69.8 gIoU and 68.2 cIoU, outperforming VideoLISA by 8.4 and 1.1 points, respectively. These results clearly demonstrate the superior performance of our model in handling both standard and reasoning-based referring image segmentation tasks.

ReVOS. The results comparison on ReVOS are shown in Table 3. VideoSeg-R1-Qwen2.5-VL-7B achieves a $\mathcal{J}\&\mathcal{F}$ score of 61.1, outperforming the best existing SFT method, ViLLa-InternVideo2-6B, by 4.1 points. Specifically, it improves $\mathcal{J}$ by 3.3 points and $\mathcal{F}$ by 4.9 points. Notably, even with only 3B parameters, VideoSeg-R1-Qwen2.5-VL-3B achieves a $\mathcal{J}\&\mathcal{F}$ of 58.6, surpassing larger models such as ViLLa (57.0) and VISA-Chat-UniVi-13B (50.9). These results verify the significant advantage of our reinforcement learning training strategy in understanding complex language expressions and reasoning video object segmentation.

Ablation Studies

SFT vs. RL. We compare the performance of SFT and RL. The baseline model is Qwen2.5-VL-7B combined with SAM2-Large. In the non-CoT setting, the thinking format reward is removed, and the model no longer generates a reasoning chain. As shown in Table 4, the SFT model performs reasonably well on in-domain data, but drops significantly on OOD datasets such as ReVOS and ReasonSeg, revealing limitations in world knowledge and multi-step reasoning. In contrast, the RL model achieves better results on both in-domain and OOD tasks, demonstrating stronger generalization. Moreover, introducing CoT rewards further improves performance. Compared to RL without CoT, RL+CoT achieves a 2.2-point gain in J&F on ReVOS and a 2.6-point improvement in gIoU on ReasonSeg, showing that the reasoning process effectively enhances the model's ability to handle OOD samples.

Dataset	$\mathcal{J}\&\mathcal{F}\uparrow$ w/	$\mathcal{J}\&\mathcal{F}\uparrow$ w/o	#Token↓ w/	#Token↓ w/o
Ref-YouTube-VOS	81.3	78.4	43.2	77.1
Ref-DAVIS17	79.8	75.2	42.8	62.4
MeViS	55.3	52.2	52.7	82.3
ReVOS	61.1	57.3	56.1	89.9

Table 7. Ablation study of Soft Length Penalty.

Hierarchical Text-guided Frame Sampler. We compare four frame selection strategies: (1) directly using the first frame (f_0); (2) directly locating the target frame; (3) randomly selecting a frame from the key segment; and (4) our proposed Hierarchical Text-guided Frame Sampler (HTFS). As shown in Table 5, HTFS achieves the best performance in both J&F and gIoU metrics, significantly outperforming the other strategies. This method simulates the human attention process that progressively shifts from coarse perception to precise focus, effectively enhancing the model's ability to locate key frames and thereby improving overall performance.

Spatial Prompts. Table 6 shows that using only Bbox significantly improves performance (J&F from 40.2 to 57.2 on ReVOS, gIoU from 56.1 to 65.7 on ReasonSeg). Adding P_{central} or P_{neg} further boosts results, highlighting their roles in localization and foreground-background discrimination. Combining all three yields the best performance (61.1 J&F, 69.8 gIoU), confirming their complementarity.

Soft Length Penalty. As shown in Table 7, introducing the Soft Length Penalty consistently improves segmentation performance across all benchmarks while significantly reducing the number of reasoning tokens. On Ref-YouTube-VOS and Ref-DAVIS17, the $\mathcal{J}\&\mathcal{F}$ scores increase by 2.9 and 4.6 respectively, with average token reductions of approximately 34 and 20. For the more challenging MeViS and ReVOS datasets, our method achieves improvements of 3.1 and 3.8 in $\mathcal{J}\&\mathcal{F}$, while reducing token usage by around 30 and 34. These results validate the effectiveness of adaptively controlling reasoning length based on task complexity.

Conclusion

We propose **VideoSeg-R1**, the first framework that introduces reinforcement learning into video reasoning segmentation. By combining hierarchical frame sampling, explicit reasoning, and decoupled segmentation-propagation, our method achieves state-of-the-art performance on multiple benchmarks. Despite strong performance, the multi-stage design and large models incur high computational cost, limiting real-time deployment. Future work will explore model simplification and tighter integration to improve practicality and scalability.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bellver, M.; Ventura, C.; Silberer, C.; Kazakos, I.; Torres, J.; and Giro-i Nieto, X. 2022. A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications*, 1–31.

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-End Referring Video Object Segmentation with Multimodal Transformers. arXiv:2111.14821.

Cheng, H. K.; and Schwing, A. G. 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, 640–658. Springer.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2694–2703. Paris, France: IEEE.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16301–16310. Montreal, QC, Canada: IEEE.

Ding, Z.; Hui, T.; Huang, J.; Wei, X.; Han, J.; and Liu, S. 2022. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4964–4973.

Du, E.; Li, X.; Jin, T.; Zhang, Z.; Li, R.; and Wang, G. 2025. GraphMaster: Automated Graph Synthesis via LLM Agents in Data-Limited Environments. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2025)*.

Du, E.; Liu, S.; and Zhang, Y. 2025a. GraphOracle: A Foundation Model for Knowledge Graph Reasoning. *arXiv preprint arXiv:2505.11125*.

Du, E.; Liu, S.; and Zhang, Y. 2025b. Mixture of Length and Pruning Experts for Knowledge Graphs Reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 432–453.

Feng, X.; Wan, Z.; Wen, M.; McAleer, S. M.; Wen, Y.; Zhang, W.; and Wang, J. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Guo, R.; Wang, X.; and Zhang, J.-F. 2019. Video Object Segmentation and Tracking: A Survey. *arXiv preprint arXiv:1904.09172*.

Guo, Y.; Lu, Y.; Zhang, W.; Xu, Z.; Chen, D.; Zhang, S.; Zhang, Y.; and Wang, R. 2025. Decoupling Continual Semantic Segmentation. arXiv:2508.05065.

He, S.; and Ding, H. 2024. Decoupling Static and Hierarchical Motion Perception for Referring Video Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13332–13341. Seattle, WA, USA: IEEE.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798. Doha, Qatar: Association for Computational Linguistics.

Kumar, A.; Zhuang, V.; Agarwal, R.; Su, Y.; Co-Reyes, J. D.; Singh, A.; Baumli, K.; Iqbal, S.; Bishop, C.; Roelofs, R.; et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Lab, E. 2025. Open R1 Multimodal. https://github.com/EvolvingLMMs-Lab/open-r1-multimodal.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9579–9589. Seattle, WA, USA: IEEE.

Lan, M.; Chen, C.; Zhou, Y.; Xu, J.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Lin, J.; Guo, Y.; Han, Y.; Hu, S.; Ni, Z.; Wang, L.; Chen, M.; Liu, H.; Chen, R.; He, Y.; Jiang, D.; Jiao, B.; Hu, C.; and Wang, H. 2025. SE-Agent: Self-Evolution Trajectory Optimization in Multi-Step Reasoning with LLM-Based Agents. arXiv:2508.02085.

Liu, C.; Ding, H.; and Jiang, X. 2023. GRES: Generalized Referring Expression Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23592–23601. Vancouver, BC, Canada: IEEE.

Liu, Y.; Qu, T.; Zhong, Z.; Peng, B.; Liu, S.; Yu, B.; and Jia, J. 2025. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning. arXiv:2505.12081.

Liu, Z.; He, Y.; Wang, W.; Wang, W.; Wang, Y.; Chen, S.; Zhang, Q.; Lai, Z.; Yang, Y.; Li, Q.; Yu, J.; Li, K.; Chen, Z.; Yang, X.; Zhu, X.; Wang, Y.; Wang, L.; Luo, P.;

Dai, J.; and Qiao, Y. 2023. InternGPT: Solving Vision-Centric Tasks by Interacting with ChatGPT Beyond Language. arXiv:2305.05662.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10031–10040. Seattle, WA, USA: IEEE.

Luo, X.; Huang, J.; Zheng, W.; Zhu, Q.; Xu, M.; Xu, Y.; Fan, Y.; Qin, L.; and Che, W. 2025. How Many Code and Test Cases Are Enough? Evaluating Test Cases Generation from a Binary-Matrix Perspective. arXiv:2510.08720.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20. Las Vegas, NV, USA: IEEE.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Gool, L. V. 2018. The 2017 DAVIS Challenge on Video Object Segmentation. arXiv:1704.00675.

R1-V Team. 2025. R1-V. https://github.com/Deep-Agent/R1-V?tab=readme-ov-file.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 3505–3506.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. PixelLM: Pixel Reasoning with Large Multimodal Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26364–26373. Seattle, WA, USA: IEEE.

Seo, S.; Lee, J.-Y.; and Han, B. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, 208–223. Springer.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Stroh, N. 2024a. TrackGPT – A generative pre-trained transformer for cross-domain entity trajectory forecasting. arXiv:2402.00066.

Stroh, N. 2024b. TrackGPT–A generative pre-trained transformer for cross-domain entity trajectory forecasting. *arXiv preprint arXiv:2402.00066*.

Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995): 476–482.

Uesato, J.; Kushman, N.; Kumar, R.; Song, F.; Siegel, N.; Wang, L.; Creswell, A.; Irving, G.; and Higgins, I. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Wang, C.; Chen, G.; Pei, B.; Yan, Z.; Zheng, R.; Xu, J.; Wang, Z.; Shi, Y.; Jiang, T.; Li, S.; Zhang, H.; Huang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2024. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. arXiv:2403.15377.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. CRIS: CLIP-Driven Referring Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11676–11685. New Orleans, LA, USA: IEEE.

Wu, D.; Wang, T.; Zhang, Y.; Zhang, X.; and Shen, J. 2023. OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE.

Wu, J.; Jiang, Y.; Bai, S.; Zhang, W.; and Bai, X. 2022a. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4984.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022b. Language as Queries for Referring Video Object Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4964–4974. New Orleans, LA, USA: IEEE.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. VISA: Reasoning Video Object Segmentation via Large Language Models. arXiv:2407.11325.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18134–18144. New Orleans, LA, USA: IEEE.

Zhang, B.; Xiao, T.; Xiao, J.; and Wei, X. 2025. Parameter-efficient weakly supervised referring video object segmentation. *Complex & Intelligent Systems*, 1–16.

Zheng, R.; Qi, L.; Chen, X.; Wang, Y.; Wang, K.; Qiao, Y.; and Zhao, H. 2025. ViLLa: Video Reasoning Segmentation with Large Language Model. arXiv:2407.14500.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023a. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15116–15127.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023b. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.