SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning

Jiaqi Huang*, **Zunnan Xu***, **Jun Zhou**†, **Ting Liu**, **Yicheng Xiao**,
Mingwen Ou, **Bowen Ji**, **Xiu Li**, **Kehong Yuan**
Tsinghua University

Abstract

Leveraging multimodal large models for image segmentation has become a prominent research direction. However, existing approaches typically rely heavily on manually annotated datasets that include explicit reasoning processes, which are costly and time-consuming to produce. Recent advances suggest that reinforcement learning (RL) can endow large models with reasoning capabilities without requiring such reasoning-annotated data. In this paper, we propose SAM-R1, a novel framework that enables multimodal large models to perform fine-grained reasoning in image understanding tasks. Our approach is the first to incorporate fine-grained segmentation settings during the training of multimodal reasoning models. By integrating task-specific, fine-grained rewards with a tailored optimization objective, we further enhance the model's reasoning and segmentation alignment. We also leverage the Segment Anything Model (SAM) as a strong and flexible reward provider to guide the learning process. With only 3k training samples, SAM-R1 achieves strong performance across multiple benchmarks, demonstrating the effectiveness of reinforcement learning in equipping multimodal models with segmentation-oriented reasoning capabilities.

1 Introduction

Multimodal Large Language Models (MLLMs) [17, 20, 37, 22, 5, 44, 51] have achieved remarkable progress in the field of visual understanding [36, 42, 27, 43, 12], with their capabilities extending to more complex and fine-grained perception tasks like multimodal segmentation [38, 21, 40]. Compared to conventional segmentation methods that rely on simple categorical labels, the reasoning segmentation task [2, 16, 31] has garnered significant attention for its flexibility and practical applicability, but it also introduces substantially greater challenges. Specifically, it requires models not only to comprehend the intent behind user-provided textual queries accurately but also to perform strong logical reasoning to generate high-quality, pixel-level segmentation outputs.

LISA [16] was the first to introduce the integration of MLLMs with segmentation models via specialized tokens, demonstrating the feasibility of applying MLLMs to reasoning segmentation tasks. Building on this foundation, subsequent studies [25, 2, 31, 41] have adopted similar strategies, leveraging task-specific tokens generated by MLLMs to improve pixel-level segmentation performance. While these approaches are promising, they often rely heavily on large-scale annotated datasets to jointly fine-tune the language model and the segmentation decoder. This not only increases training costs but also raises the risk of catastrophic forgetting, where models perform well on in-domain data but fail to generalize to out-of-domain scenarios [7]. Furthermore, the reasoning segmentation tasks frequently involve ambiguous and complex text queries from users, which demand strong reasoning

*Equal contribution
†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

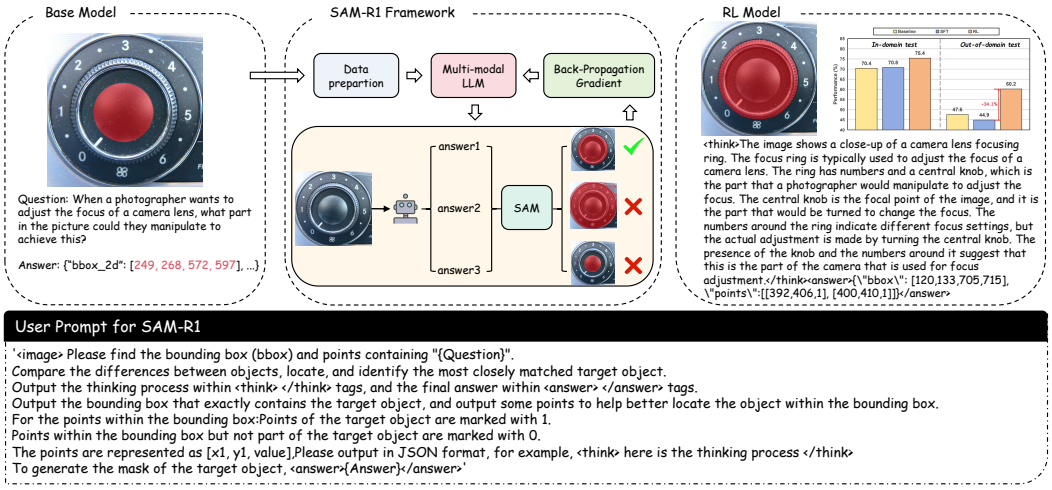


Figure 1: SAM-R1 generates a reasoning chain prior to producing the segmentation mask. It employs a reinforcement learning (RL) strategy, learning the reasoning process from scratch. In comparison to supervised fine-tuning (SFT), the RL-enhanced model, which incorporates fine-grained rewards based on SAM, demonstrates superior performance on both in-domain and out-of-domain data.

capabilities from MLLMs to accurately interpret intent and precisely localize the target segmentation regions.

Recent research has shown that reinforcement learning (RL) can significantly enhance the reasoning capabilities of large language models (LLMs) through reward-based feedback mechanisms [13]. DeepSeek-R1 [9] leverages rule-based rewards to further improve the model's capacity for complex reasoning. This method requires the model to undergo an extensive reasoning process before producing a final answer, with rewards assigned solely based on the correctness of the final response and its adherence to a predefined output format. Such rule-based reward designs align naturally with visual understanding tasks, which often come with accurate ground-truth (GT) annotations. Inspired by this, numerous efforts [8, 46, 50, 39] have applied Group Relative Policy Optimization (GRPO) [33] to vision-language models, incorporating task-specific reward signals. For example, VLM-R1 [34] introduces both format and accuracy rewards for general vision-language tasks, and further incorporates customized rewards tailored to specific applications to mitigate reward hacking. Seg-Zero [23] expands this paradigm by designing a more comprehensive reward system, including reasoning-format, segmentation-format, and accuracy rewards based on IoU and L1 distance, to stimulate robust reasoning in segmentation contexts. Although Seg-Zero demonstrates strong performance in emergent reasoning tasks, its complete decoupling of the reasoning model and segmentation decoder prevents access to pixel-level feedback from the segmentation results, thereby increasing the risk of reward hacking. To address this, involving the segmentation decoder directly in the reward loop, as a reward provider, not only ensures alignment between optimization objectives and task goals but also alleviates the need for extensive human-annotated reasoning data, enabling a more efficient and scalable learning paradigm.

Building upon the insights presented above, we propose SAM-R1, an efficient end-to-end framework tailored for reasoning segmentation. SAM-R1 utilizes reinforcement learning with reward-driven optimization to enhance the reasoning capabilities of MLLMs in complex scenarios. A key component of our framework is the design of task-specific, fine-grained reward functions, particularly a segmentation accuracy reward derived directly from the output of the Segment Anything Model (SAM). This enables the model to develop fine-grained perceptual reasoning in an end-to-end manner—an aspect that has been largely overlooked in previous multimodal reasoning models for segmentation. Integrating powerful SAM [15] has become a prevalent strategy for achieving precise pixel-level segmentation. SAM's zero-shot segmentation capabilities, facilitated by flexible prompt-based inputs, render it a highly adaptable component. While existing approaches often employ SAM as a downstream module to generate segmentation masks based on MLLM outputs [23], our framework distinguishes itself by incorporating SAM directly into the reinforcement learning training loop as a reward signal generator. This integration enables the MLLM to receive direct, task-relevant feedback

based on segmentation accuracy, thereby aligning model optimization with the final task objective in a principled and effective manner.

Moreover, we introduce a subtle modification to the clipped objective of PPO to fully utilize its potential in the reasoning segmentation task. First, we increase the upper clipping threshold to encourage updates from highly advantageous actions, thereby granting the model greater flexibility in optimizing the task-specific reasoning model. Second, we observe that GRPO may occasionally produce overly lengthy responses with limited informative content. During the GRPO optimization process, overly long responses will confuse the model and prevent it from obtaining higher reward signals, which can lead to reward hacking. Rather than constraining each token in a single response, we treat all tokens within a response group, encouraging the policy model to focus on generating responses with higher information density. By integrating task-specific, fine-grained rewards with a tailored optimization objective, SAM-R1 precisely interprets complex instructions and accurately localizes segmentation targets. Using only 3K training samples, our method surpasses the base model by 34.1% on the challenging ReasonSeg benchmark in zero-shot setting. In conclusion, our contributions can be summarized as below:

- We present a novel end-to-end framework for fine-grained, reasoning segmentation that employs rule-based rewards to enhance comprehension of complex instructions.
- We devise task-specific, fine-grained reward functions that leverage SAM as an active reward provider, driving continuous self-improvement of the reasoning model.
- We provide extensive empirical evidence demonstrating the effectiveness of SAM-R1 and offer new insights into synergizing reinforcement learning with MLLMs.

2 Related Works

2.1 MLLMs for Vision and Reasoning Segmentation

Multimodal Large Language Models (MLLMs) have significantly advanced visual understanding, extending from foundational tasks like image captioning and visual question answering [1, 26, 6] to more intricate, fine-grained perception challenges such as image segmentation. A notable direction is reasoning segmentation [45, 2, 31], which necessitates that models interpret implicit user queries and perform logical deduction to generate pixel-level masks. A relevant line of research [10, 35, 11] focuses on using a single generic prompt to perform segmentation, thereby reducing the reliance on manually provided, image-specific inputs. Seminal works like LISA [16] demonstrated the viability of MLLMs for such tasks by interfacing them with segmentation models via specialized tokens. However, these initial approaches frequently depended on Supervised Fine-Tuning (SFT) using datasets with simple categorical labels or rudimentary descriptions [23]. This reliance often curtailed out-of-domain generalization and lacked explicit, interpretable reasoning processes [23, 34], thereby motivating the exploration of methods to instill more robust reasoning capabilities within MLLMs for segmentation.

2.2 RL for Enhanced Reasoning in Multimodal Tasks

Reinforcement Learning (RL) has emerged as a potent methodology for eliciting and augmenting the reasoning capacities of large models, circumventing the need for datasets with explicit reasoning annotations. Research indicates that reward-driven optimization can effectively activate emergent test-time reasoning. Algorithms such as Group Relative Policy Optimization (GRPO) [33], employed in models like DeepSeek-R1 for language tasks [9], Seg-Zero for reasoning segmentation [23], and VLM-R1 [34] for general vision-language tasks, have achieved considerable success in training models to generate reasoning chains and attain high performance with limited supervision. These RL-based strategies often exhibit superior generalization compared to SFT methods [7], which are prone to overfitting and catastrophic forgetting of general abilities. Our work leverages this paradigm by adapting an RL training algorithm based on GRPO [33], specifically tailored to the multimodal segmentation task, to cultivate fine-grained perceptual reasoning.

2.3 Segmentation Feedback with Task-Specific Rewards

The incorporation of powerful, pre-trained segmentation models like the Segment Anything Model (SAM) [15] has become a prevalent strategy for achieving precise pixel-level segmentations. SAM's zero-shot segmentation capabilities, prompted by diverse inputs, render it a versatile component. While many frameworks employ SAM as a downstream module to produce segmentation masks from MLLM outputs [23], our approach uniquely integrates SAM as an active element within the RL training loop, functioning as a reward provider. This allows the MLLM to receive direct feedback on the quality of its generated information, assessed by the final segmentation accuracy.

The design of effective reward functions is paramount in RL. Related works often employ rule-based rewards, encompassing format rewards for structured outputs and accuracy rewards (e.g., Intersection over Union (IoU) for bounding boxes or masks, L1 distance) to quantify the quality of spatial predictions. For instance, Seg-Zero utilizes reasoning-format, segmentation format, and accuracy rewards based on IoU and L1 distance [23]. VLM-R1 also employs accuracy and format rewards for tasks such as referring expression comprehension and open-vocabulary object detection [34]. Other works like RM-R1 focus on correctness-based rewards for reward modeling itself [3], and R1-Reward introduces consistency rewards alongside formatting and result rewards for training multimodal reward models [50]. Our SAM-R1 framework is distinguished by its design of task-specific, fine-grained reward functions, notably a segmentation-accuracy reward that directly utilizes SAM's output. This enables the model to learn fine-grained reasoning for segmentation tasks in an end-to-end manner, an aspect largely overlooked in prior work on fine-grained segmentation settings within multimodal reasoning models.

3 Method

In this section, we elaborate on the architecture of our framework. In section 3.1, we explain how our framework enables multimodal large models to achieve fine-grained perceptual reasoning capacities. The enhancements made to the reinforcement learning algorithm, which significantly enhance the model's multimodal reasoning performance, are detailed in section 3.2. Furthermore, in section 3.3, we offer a detailed discussion of our approach to designing the reward function, with SAM integrated as a strong and flexible reward provider.

3.1 SAM-R1

As depicted in Figure 2, our framework takes user-supplied questions and images as input. It performs reasoning and analysis to pinpoint the target object by synthesizing information from both modalities. Subsequently, the model generates intermediate reasoning outputs, which serve as inputs to the segmentation model for mask generation. During this process, the model has the flexibility to produce outputs that enhance the segmentation model's performance. Our approach diverges from prior work [23], which centered on training the multimodal large model alone. Instead, we incorporate the segmentation model as a reward provider in the reinforcement learning phase. This integration enables the segmentation model's outputs to offer detailed feedback, thereby refining the training of the reasoning model.

3.2 RL Training Algorithm

Using reinforcement learning [9] to train large models and enhance their performance in specific domains, such as mathematics and programming, has proven effective. However, previous reinforcement learning methods often relied on a pre-trained model, which led to a significant increase in cost and complexity. At the same time, acquiring reasoning capabilities previously required carefully curated datasets that included explicit reasoning processes. Models needed to be trained on these reasoning-annotated datasets to achieve competitive performance.

Recent research [9] has shown that large models' reasoning abilities can merge even when trained on datasets without explicit reasoning rules, and the reward mechanism can be greatly simplified while still maintaining the model's strong performance.

3.2.1 DeepSeek R1-Zero and GRPO

The DeepSeek R1-Zero algorithm introduces a novel training approach using Group Relative Policy Optimization (GRPO). This method trains the model to output both a reasoning process and a final answer, while supervision is applied only to the answer. Despite this limited supervision, the model still achieves robust reasoning performance. In this framework, rule-based and accuracy-based reward functions are used to evaluate the model's responses, effectively preventing reward hacking and simplifying the overall reward mechanism.

Unlike previous reinforcement learning algorithms such as PPO [32], which require a separate critic model to evaluate performance, GRPO eliminates the need for an additional model by directly comparing all scores within a group as a baseline. Specifically, for each input question q, GRPO samples a set of G responses $\{o_1, o_2, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$. The reward advantage A_i for the i-th response is then computed by normalizing the group of rewards $\{r_1, r_2, \ldots, r_G\}$:

$$A_i = \frac{r_i - \mu_r}{\sigma_r}, \tag{1}$$

where μ_r and σ_r denote the mean and standard deviation of the rewards in the group, respectively.

Similar to PPO, GRPO adopts a clipped objective, together with a directly imposed KL penalty term:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)\right], \tag{2}$$

where the KL divergence is defined as:

$$\mathrm{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1. \tag{3}$$

3.2.2 Our Training Algorithm

Similar to recent studies [49, 24], we observe that the clipping term utilized in advantage estimation is beneficial for maintaining stability in policy updates. At the same time, the KL-divergence penalty already limits the distributional shift between successive policies and therefore also serves as a stabilizing factor. In our multimodal image-segmentation task, we aim to allow the large multimodal model greater freedom to explore finer-grained interpretations while preserving training stability. Hence, we retain the KL constraint but decouple the clipping mechanism: we replace the single

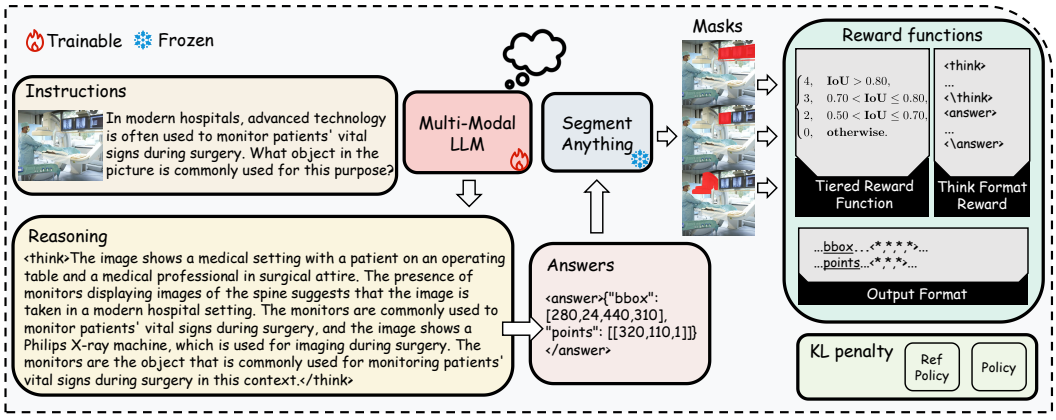


Figure 2: Our framework integrates the Segment Anything Model (SAM) as a reward provider in the reinforcement learning training of a multimodal large model (MLLM). The two models jointly process user-input questions and images to identify target objects and generate masks. Specifically, the MLLM generates the reasoning process and answer, then passes them to SAM. A fine-grained reward based on Intersection over Union (IoU) is calculated to optimize the MLLM.

threshold ε with asymmetric bounds ε_{low} and $\varepsilon_{\text{high}}$. We keep ε_{low} unchanged and slightly raise $\varepsilon_{\text{high}}$ to encourage broader exploration.

We also observe that GRPO can sometimes yield very long yet low-information answers. Such responses waste tokens and increase the risk of hallucination, as long and short answers incur the same total loss, thereby causing the per-token penalty for longer responses. To counter this, we rescale the loss so that every token receives the same loss, discouraging redundant and repetitive outputs. With these changes, our training objective becomes:

$$\begin{aligned}\mathcal{J}_{ours}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)] \\ \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\Big(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i,\right.\right. \\ \left.\left.\text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta\|\pi_{\text{ref}}\right)\Big)\right].\end{aligned} \tag{4}$$

These modifications allow the model to explore aggressively, achieve a fine-grained understanding, and train stably without incurring the extra cost and complexity of an additional critic model.

3.3 Reward Functions

A reward model is a crucial component of reinforcement learning (RL): combined with preference-alignment algorithms, it steers the policy toward the desired objectives. Following earlier work [9], we likewise employ reward functions and adapt them to the multimodal segmentation setting through three task-specific, rule-based rewards.

Tiered Segmentation-accuracy Reward Function. Departing from earlier reward designs, we treat SAM (Segment Anything Model) as an external reward provider. The target location predicted by the multimodal model is passed to SAM, which returns a mask prediction. We compute the IoU between this mask and the ground-truth mask and assign piecewise rewards:

$$\text{reward} = \begin{cases} 4, & \text{IoU} > 0.80, \\ 3, & 0.70 < \text{IoU} \leq 0.80, \\ 2, & 0.50 < \text{IoU} \leq 0.70, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

which provides robust positive feedback only when the predicted region closely aligns with the ground truth, guiding the model toward gradual improvement at lower IoU levels.

Reasoning-format reward. To encourage explicit reasoning, the model should enclose its chain-of-thought between "<think>" and "</think>" tags and place the final answer between "<answer>" and "</answer>" tags. Outputs that adhere to this structure receive a positive reward, while malformed outputs incur a penalty.

Segmentation-format reward. To ensure the multimodal large model provides fine-grained cues to the downstream segmentation module, it must emit the detected bounding box, a reference point, and a descriptive textual flag in a prescribed JSON-like format. Compliance with the schema yields a reward; deviations incur a penalty.

4 Experiment

4.1 Experimental Setup

We use Qwen2.5VL-7B [1] as our base model and SAM2-Large [29] as the segmentation model. All experiments are conducted on 8×A100 GPUs. During training, we sample 8 responses per question, set $\varepsilon_{\text{high}} = 0.3$, and use a learning rate of 1.0×10^{-6}. To ensure the model's robustness across different domains, we resize all images to 840×840 before feeding them into the MLLM during both training and evaluation. We follow previous works [16, 23] and use both cIoU and gIoU as evaluation metrics. gIoU is defined by the average of all per-image intersection-over-unions, while cIoU is defined by the cumulative intersection over the cumulative union.

4.2 Datasets

For training, we randomly sample 3,000 instances from the training set of RefCOCOg [48], which contains 104,560 referring expressions tied to 54,822 objects across 26,711 images. We use the official RefCOCOg test set as our in-domain evaluation set. To assess generalization across datasets, we use the testA subsets from RefCOCO and RefCOCO+ [14] as our out-of-distribution (OOD) evaluation sets. RefCOCO consists of 142,210 expressions for 50,000 objects across 19,994 images, while RefCOCO+ includes 141,564 expressions for 49,856 objects in 19,992 images, with both datasets providing predefined splits. RefCOCO+ is considered more challenging due to the exclusion of absolute location terms. In addition, we include ReasonSeg [16], a dataset that requires strong visual-linguistic reasoning, to further evaluate our model's ability to perform fine-grained segmentation under complex reasoning conditions.

4.3 Main Results

ReasonSeg. Table 1 shows the zero-shot performance of SAM-R1 on the ReasonSeg benchmark. Our method achieves 60.2% gIoU and 54.3% cIoU on the test set, outperforming the previous best, Seg-Zero (58.3% gIoU and 53.4% cIoU). This improvement is mainly due to our fine-grained reward design, which integrates SAM into the RL loop to provide IoU-based feedback during training, aligning reasoning with segmentation. Unlike Seg-Zero's decoupled design, our unified framework introduces finer-grained segmentation rewards, enabling stable optimization and better generalization with only 3k training samples. Additionally, our improved GRPO strategy—with asymmetric clipping and token-level loss normalization—enhances informativeness and robustness under domain shifts, supporting SAM-R1's strong zero-shot performance in complex reasoning segmentation. Seg-Zero-7B* denotes performance based on provided model weights, as their reported results used different weights per metric and could not be reproduced.

Table 1: Comparison on ReasonSeg-zero-shot benchmark (val/test). The best results are in bold.

Method	ReasonSeg-zero-shot			
	val		test	
	gIoU	cIoU	gIoU	cIoU
OVSeg [18]	28.5	18.6	26.1	20.8
ReLA [19]	22.4	19.9	21.3	22.0
Grounded-SAM [30]	26.0	14.5	21.3	16.4
LISA-7B-LLaVA1.5 [16]	53.6	52.3	48.7	48.8
LISA-13B-LLaVA1.5 [16]	57.7	60.3	53.8	50.8
SAM4MLLM [4]	46.7	48.1	-	-
Seg-Zero-7B* [23]	62.0	52.0	58.3	53.4
SAM-R1 (Ours)	**64.0**	**55.8**	**60.2**	**54.3**

Referring Expression Segmentation. Our evaluation results on the Referring Expression Segmentation datasets are shown in Table 2. We use the testA subsets of RefCOCO and RefCOCO+ as OOD test sets, and the test set of RefCOCOg as the in-domain test set. It can be seen that our model, trained on only 3,000 samples, still achieves competitive performance compared to prior methods. Specifically, on the in-domain dataset RefCOCOg, our algorithm SAM-R1 is only 0.2 points lower than Seg-Zero, despite using fewer style-consistent training samples. On the OOD datasets, our model performs comparably to Seg-Zero on RefCOCO, and improves the performance on RefCOCO+ from 73.9 to 74.7. This demonstrates the effectiveness of our approach SAM-R1. We attribute this improvement to the fine-grained reward mechanisms and the flexible exploration strategy, which allows our model to surpass previous out-of-domain performance with significantly less training data.

4.4 Visualization Analysis

As shown in Figure 3, we present some representative cases to analyze the reasoning and segmentation performance of our model in diverse scenarios.

Multiple Subjects with Fine-Grained Segmentation. In certain situations, it is necessary to identify a specific subject among multiple subjects. For example, identifying Santa Claus amidst a little girl,

Table 2: Performance comparison on referring expression benchmarks using cIoU.

Method	refCOCO	refCOCO+	refCOCOg
LAVT [47]	75.8	68.4	62.1
ReLA [19]	76.5	71.0	66.0
LISA-7B [16]	76.5	67.4	68.5
PixelLM-7B [31]	76.5	71.7	70.5
PerceptionGPT-7B [28]	78.6	73.9	71.7
Seg-Zero-7B* [23]	**79.2**	73.9	**73.3**
SAM-R1 (Ours)	**79.2**	**74.7**	73.1

chairs, Christmas trees, and various decorations, each of which is complex and numerous. The model utilizes cues, such as red clothing and the act of listening to wishes, to successfully identify and segment Santa Claus.

Global To Local Reasoning. In scenes containing rich local details, identifying a specific part from the overall structure is highly challenging. For example, in an image of an airplane composed of various components, our model accurately locates the engine by reasoning over the spatial relationship between the engine and the wings.

Challenging Environment With Distractors. In cluttered environments, such as an airport filled with various signs, identifying a specific sign, such as "Watch Your Step", poses significant challenges. Our model effectively distinguishes the target sign from visually similar ones by leveraging contextual reference objects and localizing the identification process step by step.

Complex Boundaries. For complex boundaries, such as those found in gymnastics competitions, the model integrates textual and visual information to infer that gymnastics involves specific movements. This understanding suggests the use of a vaulting table, which in turn facilitates the generation of coherent segmentation masks.

4.5 Ablation Study

In this section, we validate the effectiveness of the proposed components. As shown in Table 3, the tiered threshold strategy demonstrates superior performance compared to fixed thresholds across both in-domain and OOD benchmarks. While fixed thresholds of 0.5, 0.7, and 0.8 achieve 56.5-58.6 gIoU on the ReasonSeg-zero-shot (test), the dynamic tiered approach significantly outperforms them with 60.2 gIoU (+3.5% absolute improvement). This performance gap highlights the limitations of static thresholds in handling complex reasoning scenarios, where overly conservative predictions at high thresholds (e.g., 0.8) degrade cIoU performance despite improved localization precision. The tiered mechanism's phased threshold adjustment seems to more effectively balance precision-recall trade-offs, particularly enhancing OOD generalization, as demonstrated by its 75.4 gIoU score on the refCOCOg test set, which is 0.8% higher than the best fixed threshold.

Table 3: Ablation study on different threshold strategies under ReasonSeg-zero-shot (test) and refCOCOg-test benchmarks.

Method	ReasonSeg (gIoU)	ReasonSeg (cIoU)	refCOCOg (gIoU)	refCOCOg (cIoU)
0.5	56.5	51.9	74.7	72.8
0.7	56.2	51.6	74.9	72.6
0.8	58.6	50.8	74.6	71.9
Tiered	60.2	54.3	75.4	73.1

We further analyze the algorithmic components presented in Table 4 to validate the effectiveness of token-level constraints and the use of an asymmetric clipping strategy. The token-level reward mechanism yields consistent improvements across various metrics, enhancing ReasonSeg cIoU by 0.5% (from 51.2% to 51.7%) and refCOCOg cIoU by 0.6% (from 71.8% to 72.4%) through fine-grained output format regulation. Meanwhile, increasing the upper clipping threshold in our GRPO

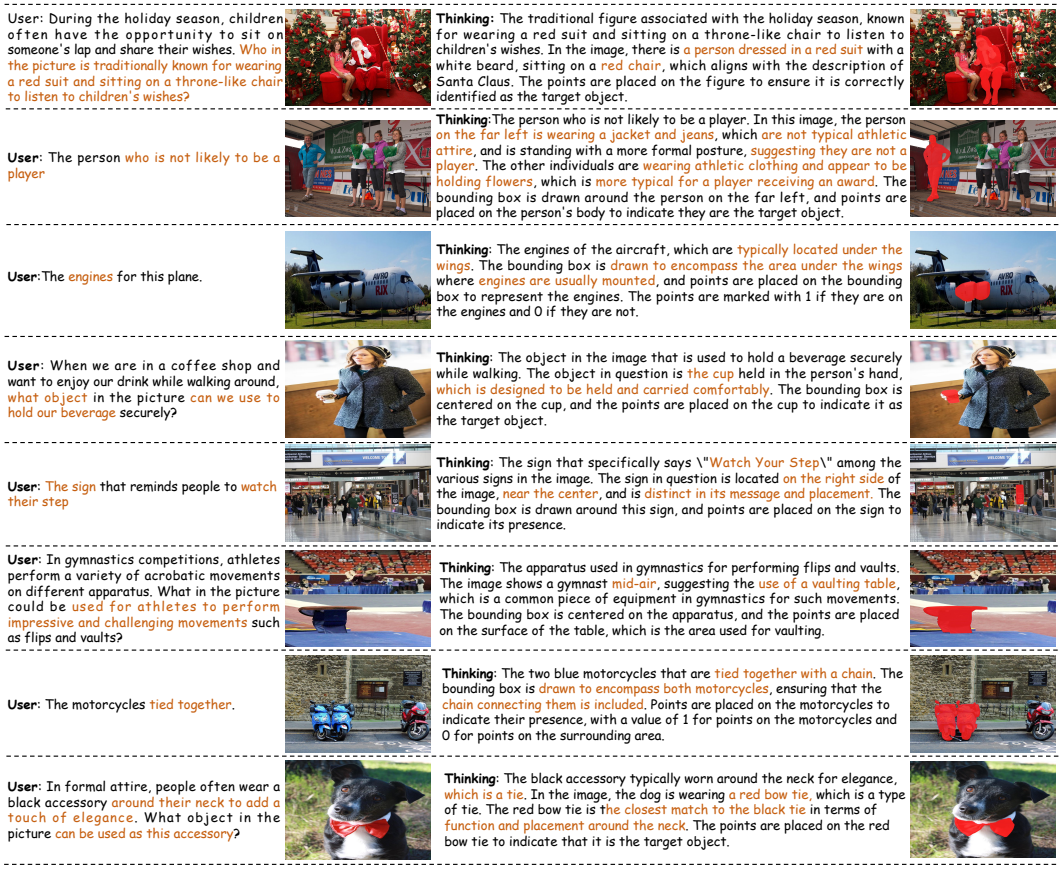


Figure 3: Qualitative results on ReasonSeg [16] demonstrate that SAM-R1 exhibits robust zero-shot performance, further enhanced by the chain-of-thought approach with improved reasoning capacity.

variant provides more flexibility in updating highly advantageous actions, which proves especially beneficial in OOD reasoning tasks. This adjustment improves ReasonSeg gIoU by 1.3%, compared to a 0.8% gain on refCOCOg, suggesting that such flexibility is more impactful in addressing complex reasoning challenges. Notably, combining both techniques yields a synergistic effect, raising ReasonSeg cIoU to 54.3%, a 3.1% improvement over the GRPO baseline. The full method achieves peak gIoU scores of 75.4 on the refCOCOg test set and 60.2 on ReasonSeg, demonstrating the effectiveness of jointly enforcing fine-grained output structure and reward-sensitive policy adaptation.

Table 4: Ablation study of algorithmic components based on the GRPO baseline on ReasonSeg-zero-shot and refCOCOg-test.

Method	Token-level	CLIP higher	gIoU (RS)	cIoU (RS)	gIoU (Rcg)	cIoU (Rcg)
GRPO	✘	✘	57.8	51.2	74.1	71.8
Token-level	✔	✘	58.0	51.7	74.5	72.4
Clip higher	✘	✔	59.1	52.8	74.9	72.5
Ours	✔	✔	60.2	54.3	75.4	73.1

4.6 Generalization to REC task

Although our model is not trained on any Referring Expression Comprehension (REC) datasets, we observe strong performance on REC task, thanks to the model's enhanced reasoning ability and fine-grained perceptual capabilities. As shown in Table 5, our method, SAM-R1, achieves state-of-the-art performance on the LISA-Grounding benchmark with 63.8, significantly surpassing previous methods such as GroundedSAM (26.2), OV-Seg (28.4), X-Decoder (28.5), and Visual-RFT (43.9).

Model	LISA-Grounding
GroundedSAM	26.2
OV-Seg	28.4
X-Decoder	28.5
Visual-RFT	43.9
SAM-R1(Ours)	**63.8**

Table 5: Performance comparison on the LISA-Grounding benchmark. Our method significantly outperforms prior open-vocabulary and vision-language segmentation approaches, demonstrating strong generalization ability on reasoning-intensive REC tasks.

This substantial improvement demonstrates the effectiveness of our reinforcement learning-based reasoning framework in complex visual grounding tasks. Unlike prior approaches, which often rely on large-scale supervised training or handcrafted prompt engineering, our method leverages task-aligned rewards and structured reasoning supervision to enable fine-grained object understanding and robust generalization in reasoning-intensive scenarios. These results demonstrate the generality and adaptability of our method beyond segmentation, highlighting its strong alignment capabilities and transferability to challenging REC scenarios.

4.7 Broader Impact and Discussion

Our work shows that reinforcement learning, guided by a segmentation model, can effectively cultivate reasoning in multimodal models. The strong performance of SAM-R1 with only 3,000 training samples highlights a promising path toward data efficiency. By using standard segmentation masks as the supervisory signal, our approach bypasses the need for costly and potentially biased, manually annotated reasoning chains, thus enhancing scalability. More broadly, this study supports a paradigm where models learn complex reasoning from task-aligned rewards rather than explicit instructions. This shift toward learning from weaker, accessible supervision is particularly impactful for domains with scarce reasoning data, such as robotic perception and medical image analysis.

We recognize several limitations for future work. First, SAM's parameters remain frozen, creating a one-way information flow that prevents it from adapting to the reasoning model. Jointly optimizing both models is a compelling next step. Though computationally demanding, this could foster a synergistic alignment where the models co-adapt. Second, our model struggles to generate meaningful negative reference points, a key capability for robust discriminative reasoning. Our RL framework failed to encourage this, suggesting a foundational limitation that may require new architectural or algorithmic solutions to improve robustness in complex visual scenes.

5 Conclusion

In this paper, we present SAM-R1, an innovative framework that leverages reinforcement learning to enhance the reasoning capabilities of multimodal large models for image segmentation. Our method introduces fine-grained segmentation settings into the training process, enabling more precise and task-relevant reasoning. Furthermore, we propose a task-specific, fine-grained reward design that incorporates the Segment Anything Model (SAM) as a flexible and reliable reward provider. By integrating these components with a tailored optimization objective, SAM-R1 achieves strong performance using only 3,000 training samples, demonstrating the practicality and effectiveness of reinforcement learning in this domain. This work not only contributes to advancing multimodal image segmentation but also highlights the potential of reward-guided learning for developing more efficient and adaptable multimodal large models.

6 Acknowledgements

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

References

[1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

[2] Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. Advances in Neural Information Processing Systems **37**, 6833–6859 (2024)

[3] Chen, X., Li, G., Wang, Z., Jin, B., Qian, C., Wang, Y., Wang, H., Zhang, Y., Zhang, D., Zhang, T., et al.: Rm-r1: Reward modeling as reasoning. arXiv preprint arXiv:2505.02387 (2025)

[4] Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)

[5] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)

[6] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)

[7] Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)

[8] Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352 (2025)

[9] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[10] Hu, J., Lin, J., Gong, S., Cai, W.: Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12511–12518 (2024)

[11] Hu, J., Lin, J., Yan, J., Gong, S.: Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. arXiv preprint arXiv:2408.15205 (2024)

[12] Huang, J., Xu, Z., Liu, T., Liu, Y., Han, H., Yuan, K., Li, X.: Densely connected parameter-efficient tuning for referring image segmentation. arXiv preprint arXiv:2501.08580 (2025)

[13] Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)

[14] Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)

[15] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

[16] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)

[17] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

[18] Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)

[19] Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: CVPR (2023)

[20] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

[21] Liu, T., Liu, X., Huang, S., Shi, L., Xu, Z., Xin, Y., Yin, Q., Liu, X.: Sparse-tuning: Adapting vision transformers with efficient fine-tuning and inference. arXiv preprint arXiv:2405.14700 (2024)

[22] Liu, T., Xu, Z., Hu, Y., Shi, L., Wang, Z., Yin, Q.: Mapper: Multimodal prior-guided parameter efficient tuning for referring expression comprehension. arXiv preprint arXiv:2409.13609 (2024)

[23] Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)

[24] Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W.S., Lin, M.: Understanding r1-zero-like training: A critical perspective. arXiv preprint arXiv:2503.20783 (2025)

[25] Luo, Z., Xiao, Y., Liu, Y., Li, S., Wang, Y., Tang, Y., Li, X., Yang, Y.: SOC: semantic-assisted object cluster for referring video object segmentation. In: NeurIPS (2023)

[26] Ma, Y., Wang, Y., Wu, Y., Lyu, Z., Chen, S., Li, X., Qiao, Y.: Visual knowledge graph for human action reasoning in videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4132–4141 (2022)

[27] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12700–12710 (2021)

[28] Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm (2023)

[29] 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., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024), `https://arxiv.org/abs/2408.00714`

[30] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)

[31] Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26374–26383 (2024)

[32] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[33] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

[34] Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)

[35] Tang, L., Jiang, P.T., Shen, Z.H., Zhang, H., Chen, J.W., Li, B.: Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 8805–8814 (2024)

[36] Wada, Y., Kaneda, K., Saito, D., Sugiura, K.: Polos: Multimodal metric learning from human feedback for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13559–13568 (2024)

[37] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

[38] Xiao, Y., Luo, Z., Liu, Y., Ma, Y., Bian, H., Ji, Y., Yang, Y., Li, X.: Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18709–18719 (2024)

[39] Xiao, Y., Song, L., Chen, Y., Luo, Y., Chen, Y., Gan, Y., Huang, W., Li, X., Qi, X., Shan, Y.: Mindomni: Unleashing reasoning generation in vision language models with rgpo. arXiv preprint arXiv:2505.13031 (2025)

[40] Xiao, Y., Song, L., Wang, J., Song, S., Ge, Y., Li, X., Shan, Y., et al.: Mambatree: Tree topology is all you need in state space model. Advances in Neural Information Processing Systems **37**, 75329–75354 (2024)

[41] Xiao, Y., Song, L., Yang, R., Cheng, C., Ge, Y., Li, X., Shan, Y.: Lora-gen: Specializing large language model via online lora generation. arXiv preprint arXiv:2506.11638 (2025)

[42] Xu, Z., Chen, Z., Zhang, Y., Song, Y., Wan, X., Li, G.: Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17503–17512 (2023)

[43] Xu, Z., Huang, J., Liu, T., Liu, Y., Han, H., Yuan, K., Li, X.: Enhancing fine-grained multi-modal alignment via adapters: a parameter-efficient training framework for referring image segmentation. In: 2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ ICML 2024) (2024)

[44] Yang, R., Song, L., Xiao, Y., Huang, R., Ge, Y., Shan, Y., Zhao, H.: Haplovl: A single-transformer baseline for multi-modal understanding. arXiv preprint arXiv:2503.14694 (2025)

[45] Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: Lisa++: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023)

[46] Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)

[47] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022)

[48] Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)

[49] Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Fan, T., Liu, G., Liu, L., Liu, X., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)

[50] Zhang, Y.F., Lu, X., Hu, X., Fu, C., Wen, B., Zhang, T., Liu, C., Jiang, K., Chen, K., Tang, K., et al.: R1-reward: Training multimodal reward model through stable reinforcement learning. arXiv preprint arXiv:2505.02835 (2025)

[51] Zhou, J., Li, J., Xu, Z., Li, H., Cheng, Y., Hong, F.T., Lin, Q., Lu, Q., Liang, X.: Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. arXiv preprint arXiv:2503.19839 (2025)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Overall, the abstract and introduction provide a concise yet comprehensive summary of the paper's objectives, methods, and findings, accurately reflecting its contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Yes, we discuss the limitations in the Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: By adhering to the principles mentioned in the Guidelines, we ensure that each theoretical result is underpinned by a full set of assumptions and complete, correct proofs, thus reinforcing the credibility and reliability of the paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the experiments section in the main text, we report all the experiment settings, implementation details, and metrics, which disclose all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be made available after being accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the experiments section in the main text, we report all the experiment settings, implementation details, and metrics, which disclose all the information needed to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: N/A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resources in the implementation details of the experiment section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we discuss both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the original owners of code, data and models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper has no crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: N/A.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

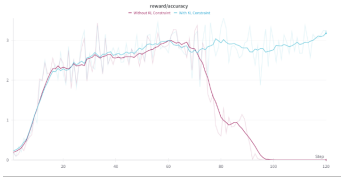
    Answer: [NA]

    Justification: The LLM is used only for formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research.

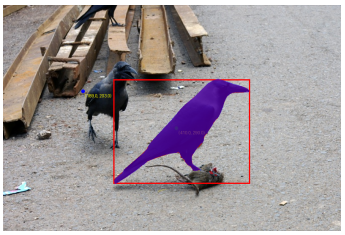
    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

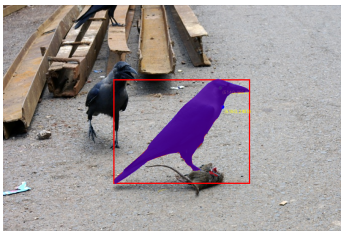
A Technical Appendices



(a) Training collapse without KL constraint.



(b) Negative points placed outside bounding boxes.



(c) All points clustered on the object when restricted to box.

Figure 4: Ablation study failures: (a) Removing the KL constraint leads to training instability and collapse. (b) Encouraging both positive and negative point generation causes negatives to appear outside target areas. (c) Forcing all points into the bounding box eliminates useful contrast, reducing performance.

A.1 Ablation Failure: Removing the KL Constraint

During the development of our method, we explored various strategies to encourage broader exploration by the model. One such attempt involved removing the KL divergence constraint, which is commonly used to regularize policy updates and limit deviation from the reference distribution.

However, empirical results showed that eliminating the KL term led to significant instability during training. As illustrated in Figure 4a, the model initially exhibited effective learning behavior with a strong exploratory signal. Yet, after approximately 100 training steps, we observed sharp fluctuations in performance, eventually leading to complete collapse of the training process.

This outcome indicates that the KL constraint plays a crucial role in maintaining training stability, especially in our multimodal reasoning setting. Consequently, we decided to retain the KL divergence term in our final framework, despite its potential to limit aggressive exploration.

A.2 Ablation Failure: Encouraging Negative Reference Points

In designing the reward function, we initially allowed the multimodal large model to freely determine the value of the reference point—positive (1) or negative (0)—without explicit supervision. However, we observed that the model strongly preferred generating only positive points, rarely including any negatives. We hypothesized that incorporating both positive and negative points could provide richer target information and improve segmentation performance.

To encourage this behavior, we introduced a format-based reward component, point value, which awarded 1 point when both 0 and 1 values appeared in the output. As shown in Figure 4b, this led the model to include both types of points. While the positive points remained well-aligned with the target object, the negative points were typically placed at the image boundaries, far outside the bounding box, offering no useful contrast for object discrimination.

We then modified the rule to grant the reward only when both positive and negative points were located within the bounding box. As shown in Figure 4c, this adjustment led to all points—regardless of label—being clustered directly on the target object, effectively eliminating the intended contrast and introducing noise instead.

These results suggest that, despite reward incentives, the multimodal large model lacks the inherent ability to identify meaningful negative examples in visual space. Therefore, we decided not to enforce negative point generation in our final design.

A.3 Failure Analysis

As illustrated in Figure 5, we present several typical failure cases, which mainly highlight two issues: incomplete segmentation and over-segmentation.

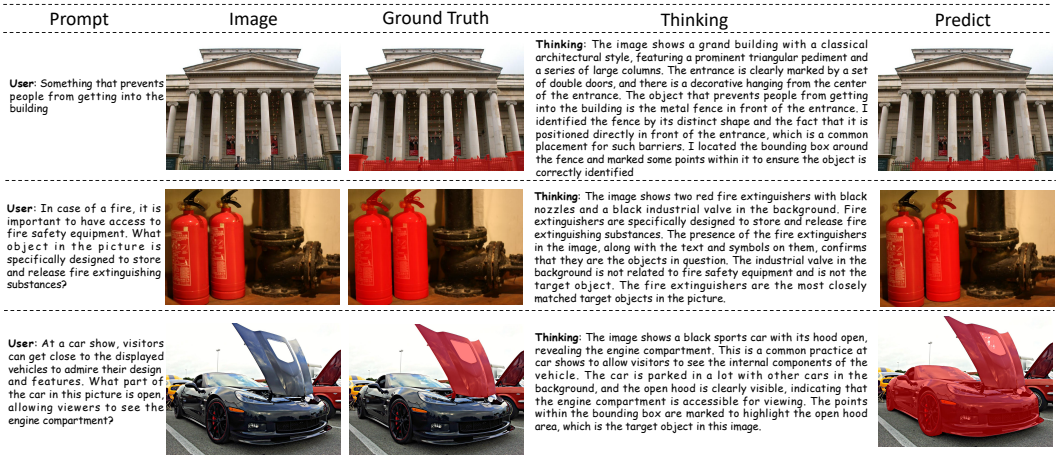


Figure 5: Visualization of some failure cases for our SAM-R1 method on the ReasonSeg-val dataset, which shows that our approach still has some limitations.

A notable observation is that our SAM-R1, through its Thinking process (fourth column), successfully comprehends the prompt and correctly localizes the target object(s). However, this correct semantic understanding does not always translate perfectly into the final segmentation mask. For example, in cases of incomplete segmentation: In the first row, the model correctly identifies the fence, but the Predict mask only covers a small portion of the target. In the second row, the model recognizes The fire extinguishers but incorrectly segments only one of the two instances. Conversely, over-segmentation is demonstrated in the third row: The prompt asks for a part of the car (the hood), and the Thinking process also pinpoints the open hood. However, the model incorrectly segments the entire vehicle instead of just the specified part.

These examples indicate that, while our model performs well in high-level semantic reasoning, limitations still exist in its ability to precisely map this understanding to pixel-level masks, particularly concerning fine-grained segmentation and instance completeness. This remains a key area for future improvement.

A.4 Data Efficiency and Scalability Analysis

To investigate the scalability and data efficiency of SAM-R1, we conducted additional experiments by increasing the size of the training data from 3k to 10k. The results clearly show that our method is highly data-efficient, with performance saturating at just 3k samples.

We present the direct comparison in Table 6. As shown, increasing the data to 10k results in negligible fluctuations in ReasonSeg: the cIoU shifts slightly from 55.8 to 55.5 on the val split and from 54.3 to 53.9 on the test split. Similarly, on the RefCOCO benchmarks, we observe only marginal gains, which strongly indicates that performance has already plateaued.

Table 6: Data efficiency analysis with 3k vs. 10k training samples.

Method	**ReasonSeg**		**RefCOCO Benchmarks**		
	Val	**Test**	**refCOCO**	**refCOCO+**	**refCOCOg**
SAM-R1 (Ours, 3k)	55.8	54.3	79.2	74.7	73.1
SAM-R1 (Ours, 10k)	55.5	53.9	79.9	75.3	73.5
Gain	-0.3	-0.4	+0.7	+0.6	+0.4

From these results, it is evident that our method's core performance saturates at 3k samples. Given the substantial increase in training cost versus the minimal performance returns, we deliberately chose 3k samples as the optimal trade-off point for demonstrating our method's capabilities.