

761 A Technical Appendices and Supplementary Material



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.

762 A.1 Ablation Failure: Removing the KL Constraint

763 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
 764 commonly used to regularize policy updates and limit deviation from the reference distribution.
 765

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

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

773 A.2 Ablation Failure: Encouraging Negative Reference Points

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

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

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

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

791 A.3 Additional Experiments: Generalization to REC task

792 Although our model is not trained on any Referring Expression Comprehension (REC) datasets, we
 793 observe strong performance on REC task, thanks to the model’s enhanced reasoning ability and
 794 fine-grained perceptual capabilities.

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.

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). 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.

A.4 Broader Impact

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. Notably, our framework avoids the need for specially curated datasets with reasoning annotations or handcrafted reasoning paths. Instead, we directly leverage standard segmentation masks as supervision, eliminating additional data preprocessing and reducing the cost of deployment in real-world scenarios.

This work contributes to the broader goal of making multimodal large models more capable and efficient for complex vision-language tasks. By showing that high-level reasoning can be acquired from weak supervision and task-aligned rewards, our findings highlight the potential for applying similar strategies to other domains, such as video understanding, robotics perception, and medical image analysis, where annotated reasoning data is scarce.

We also observe several limitations that open avenues for future improvement. Although we leverage SAM to provide fine-grained reward signals, its parameters remain frozen throughout training. Future work may explore jointly optimizing SAM alongside the multimodal model to enable deeper alignment between reasoning and segmentation. Additionally, we find that the model struggles to generate meaningful negative reference points, and our current reinforcement learning framework has limited effectiveness in enhancing this ability. Addressing this limitation may further enrich the model’s discriminative reasoning and improve robustness in complex visual scenarios.