Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning

Zuyao You
Fudan University
zyyou23@m.fudan.edu.cn

Abstract

We present **Seg-R1**, a preliminary exploration of using reinforcement learning (RL) to enhance the pixel-level understanding and reasoning capabilities of large multimodal models (LMMs). Starting with foreground segmentation tasks, specifically camouflaged object detection (COD) and salient object detection (SOD), our approach enables the LMM to generate point and bounding box prompts in the next-token fashion, which are then used to guide SAM2 in producing segmentation masks. We introduce **Group Relative Policy Optimization (GRPO)** into the segmentation domain, equipping the LMM with pixel-level comprehension through a carefully designed training strategy. Notably, Seg-R1 achieves remarkable performance with purely RL-based training, achieving **.873** S-measure on COD10K without complex model modification. Moreover, we found that pure RL training demonstrates **strong open-world generalization**. Despite being trained solely on foreground segmentation image-mask pairs without text supervision, Seg-R1 achieves impressive zero-shot performance on referring segmentation and reasoning segmentation tasks, with **71.4** cIoU on RefCOCOg test and **56.7** gIoU on ReasonSeg test, outperforming models fully supervised on these datasets. Code, weights, and datasets are available at https://geshang777.github.io/seg-r1.github.io/.

1 Introduction

Enhancing the granularity of image understanding in large multimodal models (LMMs) has long been a key challenge in the community. Early approaches have pushed the capabilities of image-level focused LMMs [26, 37, 1, 52] towards region-level [31, 59, 3, 2]. Recent research [23, 60, 45] has begun to explore ways to equip models with genuine pixel-level understanding. A common approach is to introduce special segmentation tokens into the model and design specialized decoder structures to decode these tokens into segmentation masks [23, 60]. While effective to some extent, this method disrupts the continuity of the causal architecture. Moreover, training models for segmentation via supervised fine-tuning (SFT) requires large-scale datasets with pixel-level image-text annotations and extensive training time, limiting scalability [23, 45, 60].

Recently, reinforcement learning (RL) [36, 46, 42, 47], particularly Group Relative Policy Optimization (GRPO) [47], has been widely demonstrated as more effective compared to SFT. By optimizing over groups, GRPO enables the model to explore chain-of-thought (CoT) reasoning to solve complex problems and significantly reduces memory consumption during the RLHF [38] stage. Some recent efforts [4, 48] have attempted to apply GRPO in the visual domain, proving its effectiveness in visual tasks like object counting. These studies confirm that RL, compared to SFT, can reach comparable results with substantially fewer training steps, suggesting a promising direction for efficient model training.

Inspired by these insights, we propose a new paradigm that leverages RL to equip LMMs with segmentation capabilities. We introduce **Seg-R1**, a simple yet effective framework for pixel-level learn-

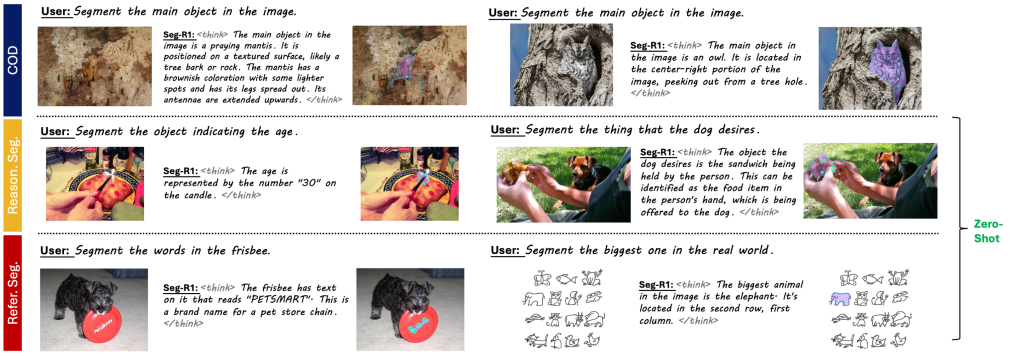


Figure 1: Overview of segmentation ability of Seg-R1.

ing. Our approach is built upon Qwen-2.5-VL [2] and SAM2 [43], where Qwen-2.5-VL is trained to generate bounding box and point prompts to guide SAM2 in producing segmentation masks. We incorporate **GRPO** into the segmentation task, requiring the model to output the reasoning process and mask prompts explicitly. To guide learning, we design a reward function that combines a format reward with a segmentation reward based on IoU and S-Measure [9], striking a balance between global accuracy and fine-grained structural fidelity.

To explore how far pure RL can drive segmentation in LMMs, we adopt a two-stage RL training strategy. Seg-R1 is first pre-trained with GRPO on the high-resolution DIS5K [41] dataset to acquire fundamental knowledge of segmentation structure and formatting. It is then further fine-tuned on COD10K [11] and CAMO [24] to enhance both its segmentation precision and reasoning ability. Notably, our method requires no architectural modifications to Qwen-2.5-VL and introduces no special tokens. Seg-R1 autonomously learns to construct annotation trajectories and generate high-quality prompts for SAM2. As a result, it achieves remarkable performance on camouflaged object detection tasks, with an S_α of .873 on COD10K-Test and 0.826 on CAMO. Besides, with further fine-tuning on DUTS [51], it achieves state-of-the-art (SoTA) performance on salient object detection, reaching an S_α of .878 on DUT-OMRON [53].

To further compare the effectiveness of RL with supervised fine-tuning (SFT), we introduce the **Foreground Chain-of-Thought (FCoT)** dataset to SFT the model as a cold start. FCoT is designed to replicate the step-by-step reasoning process a human annotator follows when using SAM2 to generate masks. It comprises $1,500$ image–mask pairs collected from existing foreground segmentation datasets. Each pair was re-annotated by fitting the foreground masks using SAM2, guided by carefully constructed bounding boxes and point prompts in accordance with a standardized annotation protocol. To explicitly capture the annotators' reasoning process, we further leverage Gemini-2.5-Pro [13] to generate a natural language chain-of-thought of the annotation steps based on the provided prompts and annotation rules. We use FCoT to fine-tune Qwen-2.5-VL via supervised learning, serving as a comparative baseline against the RL-trained version of our model.

An additional interesting finding is that, as mentioned earlier, Seg-R1 is trained solely on $7,040$ foreground segmentation image-mask pairs without any textual supervision and without exposure to referring segmentation and reasoning segmentation tasks. Despite this, it demonstrates remarkable open-world segmentation capabilities. In zero-shot evaluations on the RefCOCO [18], RefCOCO+ [56], RefCOCOg [56], and ReasonSeg [23] benchmarks, Seg-R1 achieves performance comparable to models trained directly on these datasets. Specifically, Seg-R1 attains a cIoU of 71.4 on the RefCOCOg test and 56.7 gIoU on the ReasonSeg test.

Moreover, we observe that pure RL training for segmentation does not compromise the general-purpose capabilities of the model. On general multimodal benchmarks such as MMBench [34], MME [12], POPE [27], and AI2D [19], Seg-R1 maintains performance on par with the original Qwen-2.5-VL. In contrast, the model fine-tuned via SFT experiences a noticeable performance drop. This highlights the efficacy of RL in enhancing pixel-level abilities without eroding the existing strengths of LMMs.

Our contributions can be summarized as follows:

- We propose Seg-R1, a simple yet effective RL-based framework for pixel-level comprehension in LMM.

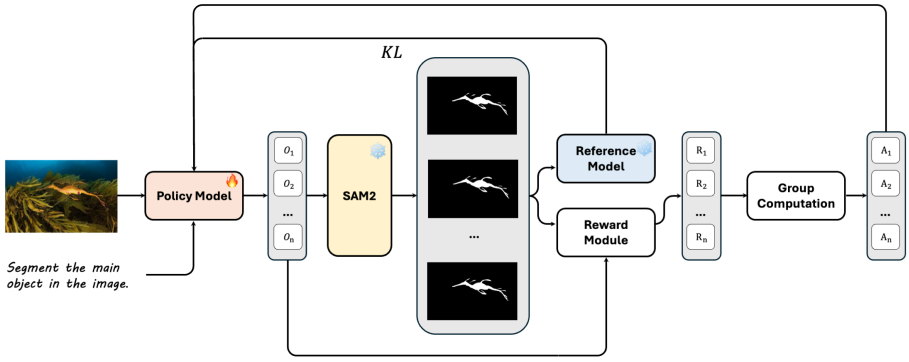


Figure 2: Overview of the Seg-R1 framework. We introduce GRPO into the segmentation domain, enabling the model to develop pixel-level understanding through group-based advantage optimization, achieving strong segmentation performance without complex architectural modifications.

- We introduce the FCoT, comprising $1,500$ manually annotated mask prompts, which provides a valuable resource for prompt-guided segmentation.
- Through comprehensive experiments, we demonstrate that pure RL training equips LMM with strong segmentation capabilities while preserving their original visual comprehension ability, outperforming SFT in terms of generalization and efficiency.

2 Related Work

Large Multimodal Models for Pixel-Level Understanding. Recent advances in LMMs have significantly enhanced the granularity of visual understanding, progressing from image-level [26, 37, 1, 52] to pixel-level comprehension [23, 60, 45]. LISA [23] first introduced the <SEG> token into LMMs, enabling LMMs with reasoning segmentation ability. Following this idea, incorporating specialized segmentation tokens to endow LMMs with pixel-level capabilities [60, 61] has become a paradigm. Despite these advances, the mixture of text tokens and segmentation tokens may lead to confusion in next-token prediction. Furthermore, SFT on pixel-level tasks can often lead to catastrophic forgetting of general-purpose capabilities [23, 45]. In contrast, we propose a novel approach that enables LMMs to perform pixel-level segmentation without complex architectural modifications, preserving both simplicity and effectiveness.

Reinforcement Learning for Large Multimodal Models. Reinforcement learning (RL) [50] has emerged as a powerful approach to enhance the training of large multimodal models (LMMs). Among the various RL techniques [36, 42, 47], Group Relative Policy Optimization (GRPO) [47] has shown particular promise. Unlike traditional RL methods, which rely on a critic model to estimate the policy model, GRPO updates the policy model from group scores, significantly reducing the computational resources required during training. Recent works [4, 48] have demonstrated the effectiveness of incorporating GRPO into LMMs, yielding substantial improvements in both visual understanding tasks. In this work, we introduce GRPO to segmentation, providing a more effective approach for improving the performance of LMMs in tasks requiring pixel-level precision.

3 Methods

3.1 Seg-R1

Segmentation has long been a challenging task in computer vision [6, 5, 20]. Training models to produce accurate masks typically requires large amounts of manually annotated pixel-level data [29, 62]. Furthermore, generating fine-grained segmentation masks usually relies on training dedicated decoder architectures that transform visual features into dense outputs [6, 5, 65], which imposes significant computational costs.

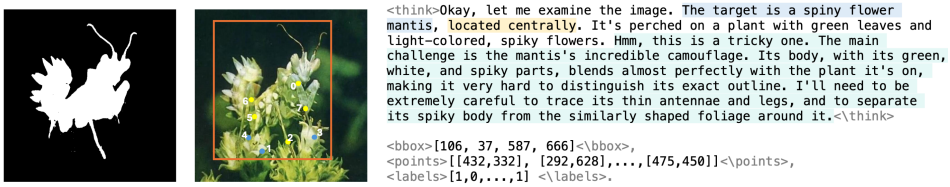


Figure 3: A visualization of FCoT, a dataset that captures the step-by-step reasoning process used by human annotators to guide SAM2 in generating segmentation masks.

In this work, we propose a much more efficient paradigm for segmentation. We employ Qwen-2.5-VL [2] to predict points, bounding boxes, and labels (referred to as mask prompts) to guide SAM2 [43] in mask generation. This approach reduces the dense prediction of segmentation to a sparse mask prompting task, significantly lowering the learning cost. Besides, using a causal LLM to predict mask prompts aligns naturally with how human annotators think and then generate masks step by step, mirroring the autoregressive nature of next-token prediction. The conditional probability of each mask prompt m_t can be formulated as $P(m_t \mid I, r, m_{<t})$, where m_t is predicted based on the input image I, the reasoning process r and the sequence of previously prompts $m_{<t}$.

We introduce GRPO into the segmentation task to enable effective reinforcement learning for the mask prompts. As illustrated in Figure 2, Given an input image and a query q, GRPO samples a group of outputs sequences $o_1, o_2, \ldots, o_G$ from the current policy $\pi_{\theta_{\text{old}}}$, and updates the policy π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{o_i \sim \pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\frac{\pi_\theta(o_i|t)}{\pi_{\theta_{old}}(o_i|t)} \hat{A}_i, \text{clip} \left(\frac{\pi_\theta(o_i|t)}{\pi_{\theta_{old}}(o_i|t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_i \right) - \beta \mathrm{D}_{KL}(\pi_\theta || \pi_{ref}) \right] \quad (1)$$

Here, $\hat{A}_i$ represents the estimated advantage, which measures the relative quality of an output within a group. The policy update is clipped by a threshold ϵ to ensure stable training. To prevent the updated policy from drifting too far from the original behavior, a KL divergence penalty is applied between the π_θ and the reference model π_{ref}, scaled by a hyper-parameter β.

The reward module is a combination of the format reward and the segmentation reward. The format reward is assigned a value of 1.0 if the output strictly wraps the generated content within the designated tags (<think></think>, <bbox></bbox>, <points></points>, and <labels></labels>); otherwise, the format reward is 0. We initially use S-Measure [9] as the segmentation reward. S-measure combines both structural and content similarity between the predicted and ground truth, better aligns with human perceptual understanding by capturing both global structure and local consistency. However, using S-measure alone leads to reward hacking as mentioned in Sec 4.4. To mitigate this, we use a weighted combination of IoU (0.7) and S-measure (0.3) as the segmentation reward, promoting both global accuracy and structural consistency.

3.2 FCoT

To further investigate the differences between supervised fine-tuning (SFT) and reinforcement learning (RL) in the segmentation domain, we require a dataset that supports structured prompt-based supervision, specifically, one that incorporates bounding boxes and point-based prompts to guide SAM2 in mask generation. However, such datasets are currently lacking. To fill this gap, we introduce **FCoT**, a dataset specifically designed to capture the step-by-step reasoning process that human annotators follow when using SAM2 to produce segmentation masks.

In practice, human annotators interacting with SAM2 follow a standardized multi-step procedure: (1) identifying and locating the target object, (2) grounding the object using a bounding box, and (3) refining the segmentation using a combination of foreground and background points. During the annotation process for FCoT, we adhere strictly to this protocol and carefully record all points and bounding box prompts. This enables the model to explicitly learn from the reasoning trajectory embedded in human annotation behavior. To enrich the dataset with interpretable reasoning paths, we

Table 1: Results on Camouflaged Object Detection (COD). ⋄ indicates SFT on FCoT as a cold start.

	COD10K [11]				CAMO [24]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
Mask Supervision Setting								
PFNet [35]	.800	.660	.868	.040	.782	.695	.852	.085
ZoomNet [39]	.838	.729	.911	.029	.820	.752	.892	.066
BSA-Net [64]	.818	.699	.901	.034	.794	.717	.867	.079
FSPNet [17]	.851	.735	.895	.026	.856	.799	.899	.050
ZoomNeXT [40]	.898	.827	.956	.018	.889	.857	.945	.041
FOCUS [55]	**.910**	**.883**	**.974**	**.013**	**.912**	**.904**	**.963**	**.025**
Prompt-Guided Setting								
SAM[21]	.730	.673	.737	.093	.643	.597	.639	.160
GPT4V+SAM [37, 21]	.601	.448	.672	.187	.573	.466	.666	.206
GenSAM [15]	.783	.695	.843	.058	.729	.669	.798	.106
ProMaC [16]	.805	.716	.876	.042	.767	.725	.846	.090
Grounded SAM2 [44]	.686	.628	.722	.209	.578	.571	.612	.302
Seg-R1-3B⋄	.850	.798	.902	.036	.810	**.798**	.861	.077
Seg-R1-3B	.857	.816	.908	.033	.805	.797	.853	.079
Seg-R1-7B	**.873**	**.820**	**.926**	**.031**	**.826**	.788	**.881**	**.073**

Table 2: Results on Salient Object Detection (SOD). ⋄ indicates SFT on FCoT as a cold start, **ft** refers the version fine-tune on the DUTS-TR.

	DUT-OMRON [53]				DUTS-TE [51]				HKU-IS [25]				ECSSD [49]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
Mask Supervision Setting																
EVPv2 [33]	.862	.857	.895	.047	.915	.923	.948	.027	.932	.953	.963	.023	.935	.958	.957	.028
SelfReformer [58]	.859	.838	.884	.043	.911	.916	.920	.026	.930	.947	.959	.024	.941	**.963**	.935	.025
FOCUS [55]	**.868**	.836	**.900**	.045	**.929**	**.928**	**.965**	**.019**	**.935**	.942	**.974**	**.018**	**.943**	.954	**.971**	**.018**
Weakly Supervised Setting																
HSS [7]	–	–	–	.050	.837	.807	–	.050	.887	.892	–	.038	.886	.899	–	.051
A2S [63]	.719	.841	.069	–	.750	.860	.065	–	.887	.937	.042	–	.888	.911	.064	–
A2SV3 [57]	–	.759	.868	.062	–	.816	.906	.047	–	.908	.954	.033	–	.923	.951	.038
Prompt-Guided Setting																
Grounded SAM2 [44]	.708	.590	.744	.102	.800	.748	.833	.078	.825	.842	.864	.069	.845	.849	.871	.078
Seg-R1-3B⋄	.837	.789	.878	.048	.861	.848	.856	.048	.780	.818	.820	.077	.881	.871	.901	.052
Seg-R1-3B	.852	.821	.890	.045	.866	.863	.899	.045	.772	.814	.809	.078	.882	.908	.903	.050
Seg-R1-3B (ft)	.868	.839	.906	.047	.909	.914	.942	.031	.890	.914	.925	.041	.916	.939	.940	.036
Seg-R1-7B (ft)	**.878**	**.850**	**.911**	**.045**	**.925**	**.922**	**.953**	**.025**	**.935**	**.950**	**.966**	**.022**	**.939**	**.956**	**.962**	**.025**

utilized Gemini-2.5-Pro [13] to generate natural language explanations based on the image, recorded annotation steps, and the annotation rules.

FCoT comprises 1,500 image–mask pairs curated from existing foreground segmentation datasets: $1,000$ images from DUTS [51], 400 from COD10K [11], and 100 from CAMO [24]. Each mask was re-annotated by replacing the original dense mask with structured mask prompt sequences and corresponding chain-of-thought annotations. We provide a visualization of FCoT in Figure 3.

4 Experiments

4.1 Training Strategies and Implementation Details

We explore two training paradigms for Seg-R1: SFT followed by RL, and pure RL from scratch.

SFT Followed by RL. To provide Seg-R1 with a solid initialization, we perform supervised fine-tuning (SFT) on the FCoT for one epoch as a cold start. This step equips the model with a basic understanding of output format and prompt-guided segmentation. During SFT, we use a learning rate of 2×10^{-5} and a batch size of 128. Following this stage, we apply reinforcement learning using the GRPO [47] algorithm. The model is trained exclusively on the camouflaged object detection dataset COD10K [11] and CAMO [24], with four samples generated per prompt during policy optimization. To improve efficiency and memory usage, we employ vLLM [22] with Flash Attention V2 [8]. During the RL stage, the learning rate is set to 10^{-6} with a batch size of 24.

Pure RL Training. To investigate the full potential of reinforcement learning in pixel-level tasks, we also train Seg-R1 entirely from scratch using RL. We begin with pre-RL training on the DIS5K-TR [41] dataset, which consists of $3,000$ high-resolution images with meticulously annotated masks. Compared to low-resolution datasets, DIS5K provides richer structural details, encouraging more nuanced model reasoning. We only require the model to generate points and labels as mask prompts

Figure 4: Comparison of single object referring segmentation in the wild.

in this stage. After pre-RL, we further RL the model on COD10K and CAMO. We use the same RL recipe as in the previous paradigm. Images are resized to 768×768 for training and 1024×1024 for inference. All experiments are conducted on 8 NVIDIA A100 GPUs with 80G memory.

4.2 Benchmark and Metrics

Foreground segmentation. We evaluate Seg-R1 on two representative tasks: camouflaged object detection (COD) and salient object detection (SOD). For COD, we adopt widely used CAMO [24] and COD10K [11] as benchmarks. And for SOD, we use DUTS-TE [51], DUT-OMRON [53], HKU-IS [25], and ECSSD [49] as benchmarks. We use S-measure (S_α) [9], E-measure (E_ϕ) [10], F-measure (F_β) (weighted F-Measure for COD and max F-Measure for SOD), and Mean Absolute Error (M) as metrics, which are commonly used in foreground segmentation to comprehensively evaluate the models.

Referring segmentation and reasoning segmentation. For referring segmentation, we evaluate on RefCOCO [18], RefCOCO+ [56], and RefCOCOg [56], using cIoU as the metric. For reasoning segmentation, we conduct experiments on the ReasonSeg [23] dataset and report both cIoU and gIoU to capture the capability of the model in open-world pixel-level understanding.

4.3 Comparing with State-of-the-arts

Results on COD. COD focuses on segmenting disguised objects that blend seamlessly into their surroundings, *e.g.*, mimetic organisms. This task is particularly challenging due to the low visual saliency and ambiguous object boundaries. We compare the recently proposed methods that directly supervise the mask or supervise prompts (*i.e.*, points, bounding box, and scribble) in COD. As shown in the Table 1, Seg-R1 consistently outperforms the prompt supervised methods, surpassing strong baselines such as Grounding SAM2 [44], ProMaC [14], and GenSAM [15] by a significant margin. But, considering that some of these prompt-guided methods are weakly supervised, we also compare Seg-R1 with previous fully supervised methods. However, the segmentation capability of Seg-R1 relies heavily on frozen SAM2, which is known to struggle with camouflaged objects [15]. As a result, there remains a performance gap on CAMO between Seg-R1 and prior fully supervised models specifically designed for camouflage segmentation. Nevertheless, our model still achieves competitive scores on COD10K, with an S_α of .873 on COD10K-Test.

Results on SOD. SOD aims to segment the most salient part of the image from the background. We compare the recent SOD methods on four widely used benchmarks in the Table 2. We first evaluate the zero-shot performance on SOD, with the version RL only on the DIS5K, COD10K, and CAMO datasets. As shown in the table, Seg-R1 demonstrates comparable performance with previous SoTA, which is supervised on DUTS datasets. We further fine-tune Seg-R1 on the training set of DUTS and find that Seg-R1 achieves SoTA performance in the salient object detection benchmarks. Notably, Seg-R1 achieves an S_α of .878 and an E_ϕ of .911 on DUT-OMRON [53], outperforming the previous mask or prompt supervised methods with a clear margin.

Zero-Shot Transfer. Since reinforcement learning enables the model to acquire human-like prompting strategies for segmenting target objects, we further investigate whether this capability generalizes to other tasks. Specifically, we explore referring segmentation and reasoning segmentation—two

Table 3: Results on referring segmentation.

	RefCOCO			RefCOCO+			RefCOCOg	
	testA	testB	val	testA	testB	val	test	val
LAVT [54]	75.8	68.8	72.7	68.4	55.1	62.1	66.0	65.0
X-Decoder [65]	–	–	–	–	–	–	–	64.6
SEEM [66]	–	–	–	–	–	–	–	65.7
LISA-7B [23]	76.5	71.1	74.1	67.5	56.5	62.4	68.5	66.4
PixelLM-7B [45]	76.5	68.2	73.0	**71.7**	58.3	**66.3**	70.5	69.3
OMG-LLaVA-7B [60]	77.7	**71.2**	**75.6**	69.7	**58.9**	65.6	70.2	70.7
Seg-R1-3B◇	65.8	54.7	58.7	56.2	45.0	49.1	57.0	57.9
Seg-R1-3B	76.0	64.9	69.9	66.8	50.9	59.1	67.9	67.3
Seg-R1-7B	**78.7**	67.6	74.3	70.9	57.9	62.6	**71.4**	**71.0**

Table 4: Results on ReasonSeg.

	val		test	
	gIoU	cIoU	gIoU	cIoU
OVSeg [28]	28.5	18.6	26.1	20.8
GRES [30]	22.4	19.9	21.3	22.0
X-Decoder [65]	22.6	17.9	21.7	16.3
SEEM [66]	25.5	21.2	24.3	18.7
Grounded-SAM [32]	26.0	14.5	21.3	16.4
LISA-7B [23]	44.4	46.0	36.8	34.1
Seg-R1-3B◇	50.3	35.5	42.4	30.0
Seg-R1-3B	**60.8**	**56.2**	55.3	46.6
Seg-R1-7B	58.6	41.2	**56.7**	**53.7**

Table 5: Ablation on training strategy.

	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
baseline	.695	.682	.734	.181
$w.$ RL	.724	.691	.749	.127
$w.$ SFT	.790	.770	.841	.093
$w.$ SFT + RL	.810	.798	.861	.077
$w.$ Pre-RL + RL	.805	.787	.853	.079

Table 6: Ablation on segmentation reward function.

	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
baseline	.614	.532	.439	.151
$w.$ IoU	.785	.770	.889	.084
$w.$ S-Measure	.410	.049	.251	.181
$w.$ IoU + S-Measure	.805	.787	.853	.079

challenging benchmarks that demand both visual and language understanding. Referring segmentation requires identifying and segmenting specific objects based on natural language expressions, while reasoning segmentation involves interpreting complex, multi-step instructions or contextual logic to produce accurate segmentation masks.

Surprisingly, we find that our model, trained solely via reinforcement learning on $7,040$ foreground segmentation image-mask pairs without any textual supervision, exhibits strong zero-shot generalization to both tasks. On RefCOCOg [56], Seg-R1 achieves 71.4 and 71.0 on the validation and test sets, comparable with the performance of current SoTA models. On the test set of ReasonSeg [23], it reaches 56.7 in gIoU, significantly outperforming the previous models like LISA-7B [23].

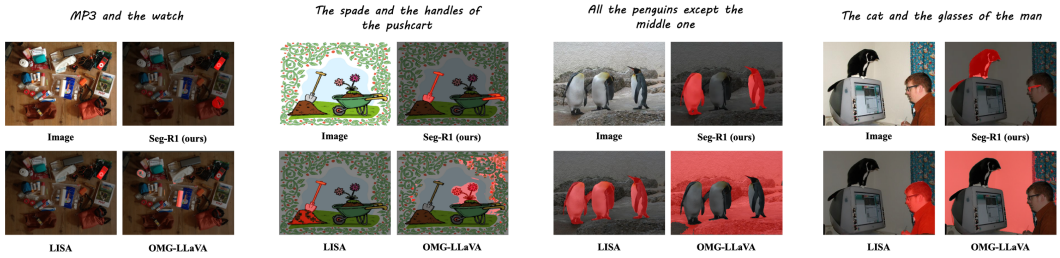


Figure 5: Comparison of multiple objects referring segmentation in the wild.

While previous models achieve high scores on benchmarks like RefCOCO, our evaluation reveals a significant gap between benchmark performance and real-world applicability. As shown in Figure 4 and Figure 5, we test the referring segmentation ability of the model in the wild and observe that existing models such as LISA [23] and OMG-LLaVA [60] fail on many cases involving real-world complexity, such as sketches and fine-grained segmentation. In contrast, Seg-R1 consistently demonstrates accurate segmentation across a wide range of open-vocabulary expressions. Whether in real-world images, hand-drawn illustrations, or multi-object scenarios, Seg-R1 shows remarkable generalization and robustness in understanding and grounding expressions for segmentation.

4.4 Ablation Study

Effects of training strategy. We compare the impact of different training strategies on model performance, as summarized in Table 5. We use Qwen-2.5-VL-3B as our baseline and CAMO as the benchmark. Specifically, we prompt Qwen-2.5-VL to generate a bounding box for the target object, and then use it to guide SAM2 for mask generation. Our experiments show that both supervised fine-tuning (SFT) and reinforcement learning (RL) significantly enhance segmentation performance. Notably, SFT with the proposed FCoT dataset as cold start or pre-training with the high-resolution

Figure 6: Visualization of segmentation reward curve.

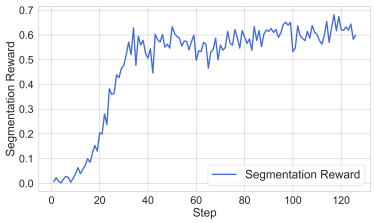
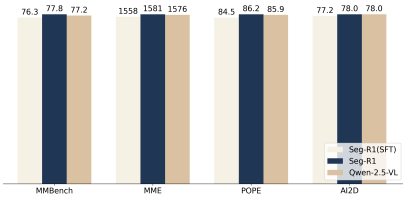


Figure 7: Comparison of the general VLM benchmarks.



DIS5K dataset prior to RL both lead to substantial gains on the COD task. Building upon these improvements, further applying RL on COD10K brings additional benefits, consistently boosting segmentation performance. Besides, we observe a promising reward curve during the RL process as shown in Figure 6.

Effects of Reward Function. We explore the impact of different reward functions during GRPO training. The model pre-RL on DIS5K serves as our baseline, and CAMO serves as the benchmark. As shown in Table 6, using only IoU as the reward leads to significant performance improvements on the CAMO benchmark. However, relying solely on IoU may cause the model to overlook structural details within the mask. To address this, we also experimented with using only the S-Measure as the reward. The results show that this leads to reward hacking—without IoU as a corrective signal, the model tends to predict entirely black masks, resulting in poor performance on several metrics. To mitigate this issue, we adopt a combined reward function that incorporates both IoU and S-Measure. As demonstrated in the table, this combination consistently achieves the best results in COD tasks, effectively balancing overall mask accuracy with structural fidelity.

Comparison of RL and SFT. As shown in Tables 1 and 2, pure RL strategy and SFT followed by RL yield comparable performance on foreground segmentation tasks. However, the pure RL strategy exhibits far superior generalization capabilities compared to the SFT-based method. As illustrated in Tables 3 and 4, we evaluate both models on two out-of-domain tasks: referring segmentation and reasoning segmentation. We find that the RL-trained model successfully transfers its learned segmentation skills from foreground segmentation to these new domains by leveraging the strong visual reasoning and grounding capabilities of Qwen-2.5-VL. In contrast, the SFT-based model performs relatively poorly on both tasks.

Moreover, we further evaluate the models on a range of general multimodal benchmarks. As shown in Figure 7, the pure RL version of Seg-R1 retains performance on par with the original Qwen-2.5-VL across MMBench [34], MME [12], POPE [27], and AI2D [19]. In contrast, the SFT version suffers from performance degradation, indicating a loss of general capabilities.

These findings highlight a key advantage of the RL approach: it equips the model with strong segmentation capabilities without sacrificing its original reasoning and understanding abilities, whereas SFT tends to overfit to the segmentation task at the expense of general multimodal performance.

5 Conclusion

In this work, we propose **Seg-R1**, a simple yet highly effective paradigm for enabling LMMs to perform segmentation. Instead of relying on dense mask generation or complex architectural modifications, we reformulate segmentation as a next-mask-prompt prediction task, significantly reducing the computation cost and aligning naturally with the causal structure of autoregressive models. Through extensive experiments, we demonstrate that: (1) Pure RL can endow LMMs with powerful segmentation abilities. Seg-R1 achieves SoTA performance on multiple segmentation benchmarks; (2) The segmentation ability acquired via RL generalizes well to out-of-domain scenarios. Trained solely on image-mask pairs without text supervision, Seg-R1 exhibits impressive zero-shot performance on both referring segmentation and reasoning segmentation tasks, even under open-world settings; (3) RL preserves the original general-purpose capabilities of the LMM, whereas SFT tends to compromise performance on general visual comprehension.

Despite these promising results, there remains room for future exploration. For instance, further improve the performance of Seg-R1 on COD and extend support to multi-turn interaction. We plan to explore these in the future.

References

[1] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[3] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[4] L. Chen, L. Li, H. Zhao, Y. Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025. Accessed: 2025-02-02.

[5] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. *CVPR*, 2022.

[6] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021.

[7] R. Cong, Q. Qin, C. Zhang, Q. Jiang, S. Wang, Y. Zhao, and S. Kwong. A weakly supervised learning framework for salient object detection via hybrid labels. *TCSVT*, 2022.

[8] T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[9] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[10] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

[11] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *CVPR*, 2020.

[12] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[13] Google DeepMind. Gemini 2.5 technical report. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf, 2025. Accessed: 2025-06-18.

[14] J. Hu, Z. Cheng, and S. Gong. Int: Instance-specific negative mining for task-generic promptable segmentation. *arXiv preprint arXiv:2501.18753*, 2025.

[15] J. Hu, J. Lin, S. Gong, and W. Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *AAAI*, 2024.

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

[17] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, 2023.

[18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[19] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[20] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *CVPR*, 2019.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *ICCV*, 2023.

[22] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, 2023.

[23] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

[24] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[25] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.

[26] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[27] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[28] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[30] C. Liu, H. Ding, and X. Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, 2023.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *NeurIPS*, 2023.

[32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[33] W. Liu, X. Shen, C.-M. Pun, and X. Cun. Explicit visual prompting for universal foreground segmentations. *arXiv preprint arXiv:2305.18476*, 2023.

[34] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

[35] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021.

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 2015.

[37] OpenAI. Gpt-4 technical report, 2023.

[38] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[39] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, 2022.

[40] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *TPAMI*, 2024.

[41] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. V. Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

[42] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

[43] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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

[45] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, pages 26374–26383, 2024.

[46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

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

[49] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 2015.

[50] R. S. Sutton, A. G. Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[51] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[52] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[53] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[54] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[55] Z. You, L. Kong, L. Meng, and Z. Wu. FOCUS: Towards universal foreground segmentation. In *AAAI*, 2025.

[56] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[57] Y. Yuan, W. Liu, P. Gao, Q. Dai, and J. Qin. Unified unsupervised salient object detection via knowledge transfer. *arXiv preprint arXiv:2404.14759*, 2024.

[58] Y. K. Yun and W. Lin. Towards a complete and detail-preserved salient object detection. *TMM*, 2023.

[59] S. Zhang, P. Sun, S. Chen, M. Xiao, W. Shao, W. Zhang, Y. Liu, K. Chen, and P. Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *ECCV*, 2025.

[60] T. Zhang, X. Li, H. Fei, H. Yuan, S. Wu, S. Ji, C. C. Loy, and S. Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *NeurIPS*, 2024.

[61] T. Zhang, X. Li, Z. Huang, Y. Li, W. Lei, X. Deng, S. Chen, S. Ji, and J. Feng. Pixel-sail: Single transformer for pixel-grounded understanding. *arXiv preprint arXiv:2504.10465*, 2025.

[62] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[63] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *CVPR*, 2023.

[64] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. In *AAAI*, 2022.

[65] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023.

[66] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *NeurIPS*, 2023.