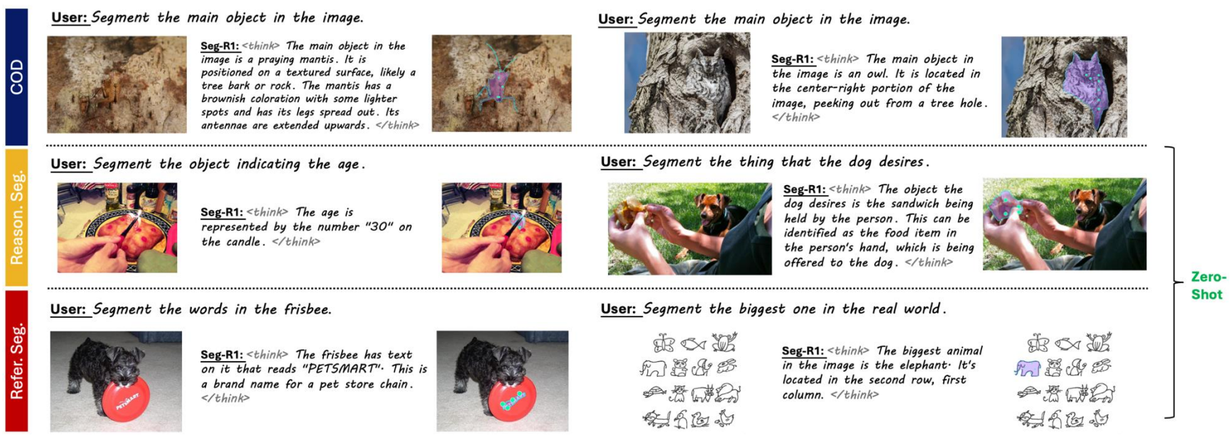
Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning

Project Page | Source Code | Web Demo

have fun :)

Zuyao You[1]
[1]Fudan University

Introduction



TL;DR: In this paper, we explore how to endow large language models (LLMs) with open-world segmentation capabilities using **purely reinforcement learning**, relying solely on foreground segmentation data **without any textual labels**.

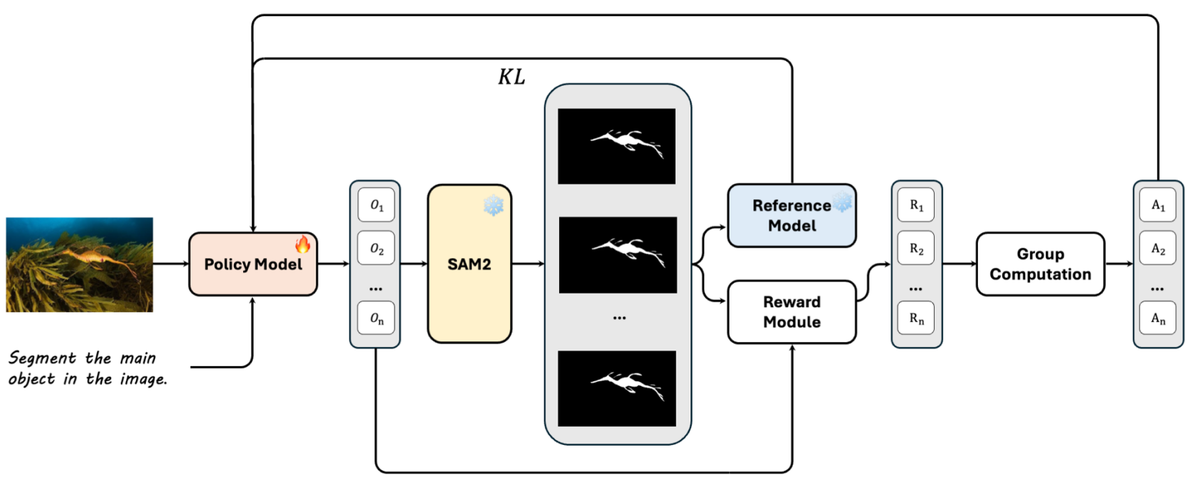
Large multimodal models (LMMs) have made rapid progress in image understanding, yet their pixel-level comprehension remains limited. Existing approaches typically introduce special segmentation tokens or rely on large-scale supervised fine-tuning with dense annotations, which breaks architectural simplicity and restricts scalability.

Key Contributions

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

Method

GRPO for Segmentation



We introduce Seg-R1, a new RL-driven paradigm that equips large multimodal models with pixel-level segmentation capabilities. we adopt a two-stage RL training strategy using GRPO, guiding the model to output explicit reasoning steps and mask prompts, while a reward function combining format constraints, IoU, and S-Measure ensures both global accuracy and fine-grained structural fidelity.

FCoT


```
<think>Okay, let me examine the image. The target is a spiny flower
mantis, located centrally. It's perched on a plant with green leaves and
light-colored, spiky flowers. Hmm, this is a tricky one. The main
challenge is the mantis's incredible camouflage. Its body, with its green,
white, and spiky parts, blends almost perfectly with the plant it's on,
making it very hard to distinguish its exact outline. I'll need to be
extremely careful to trace its thin antennae and legs, and to separate
its spiky body from the similarly shaped foliage around it.<\think>

<bbox>[106, 37, 587, 666]<\bbox>,
<points>[[432,332], [292,628],...,[475,450]]<\points>,
<labels>[1,0,...,1] <\labels>.
```

To compare RL with supervised fine-tuning (SFT), we construct the **Foreground Chain-of-Thought (FCoT)** dataset as a cold-start resource for SFT. FCoT contains 1,500 re-annotated image–mask pairs, each fitted using SAM2 with standardized mask prompts and step-by-step thinking process.

Experiment

In-domain Tasks

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**

Out-of-domain Tasks

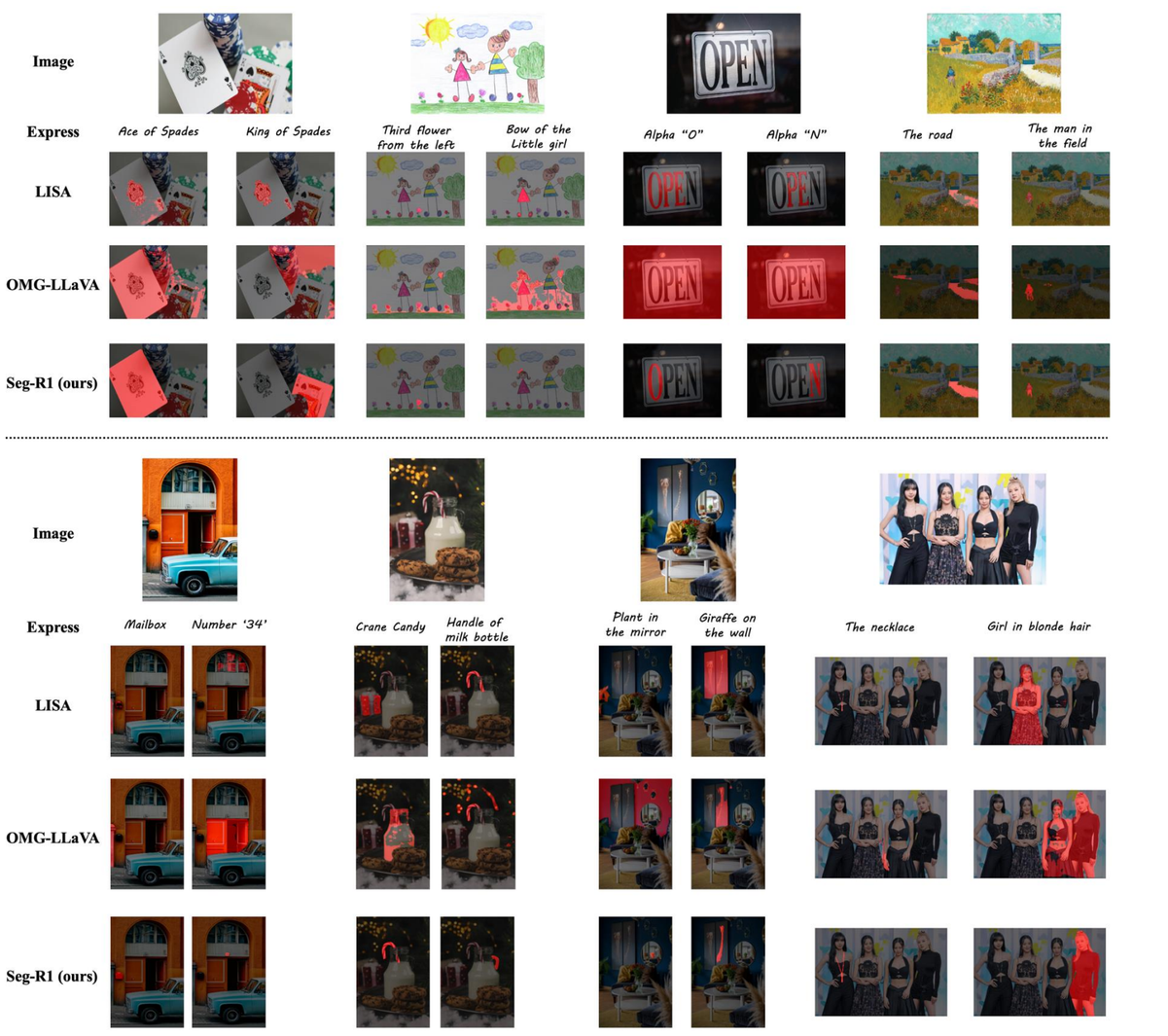
Table 3: Results on referring segmentation.

	RefCOCO			RefCOCO+			RefCOCOg	
	testA	testB	val	testA	testB	val	test	val
LAVT [63]	75.8	68.8	72.7	68.4	55.1	62.1	66.0	65.0
X-Decoder [78]	–	–	–	–	–	–	–	64.6
SEEM [79]	–	–	–	–	–	–	–	65.7
LISA-7B [24]	76.5	71.1	74.1	67.5	56.5	62.4	68.5	66.4
PixelLM-7B [53]	76.5	68.2	73.0	**71.7**	58.3	**66.3**	70.5	69.3
OMG-LLaVA-7B [71]	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.

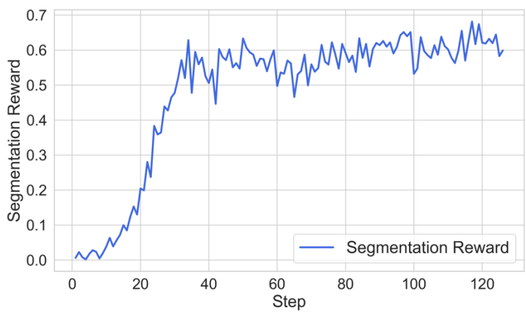
	val		test	
	gIoU	cIoU	gIoU	cIoU
OVSeg [30]	28.5	18.6	26.1	20.8
GRES [32]	22.4	19.9	21.3	22.0
X-Decoder [78]	22.6	17.9	21.7	16.3
SEEM [79]	25.5	21.2	24.3	18.7
Grounded-SAM [34]	26.0	14.5	21.3	16.4
LISA-7B [24]	44.4	46.0	36.8	34.1
Seg-R1-3B◇	50.3	35.5	42.4	30.0
Seg-R1-3B	**60.8**	**56.2**	55.3	46.6
Seg-R1-7B	58.6	41.2	**56.7**	**53.7**

Single object referring segmentation in the wild



Multiple objects referring segmentation in the wild

- **The training curve of segmentation reward.**



- **Comparison on General VLM Benchmarks**