Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement

Yuqi Liu[1*] Bohao Peng[1*] Zhisheng Zhong[1] Zihao Yue[3] Fanbin Lu[1] Bei Yu[1] Jiaya Jia[2]

CUHK[1] HKUST[2] RUC[3]

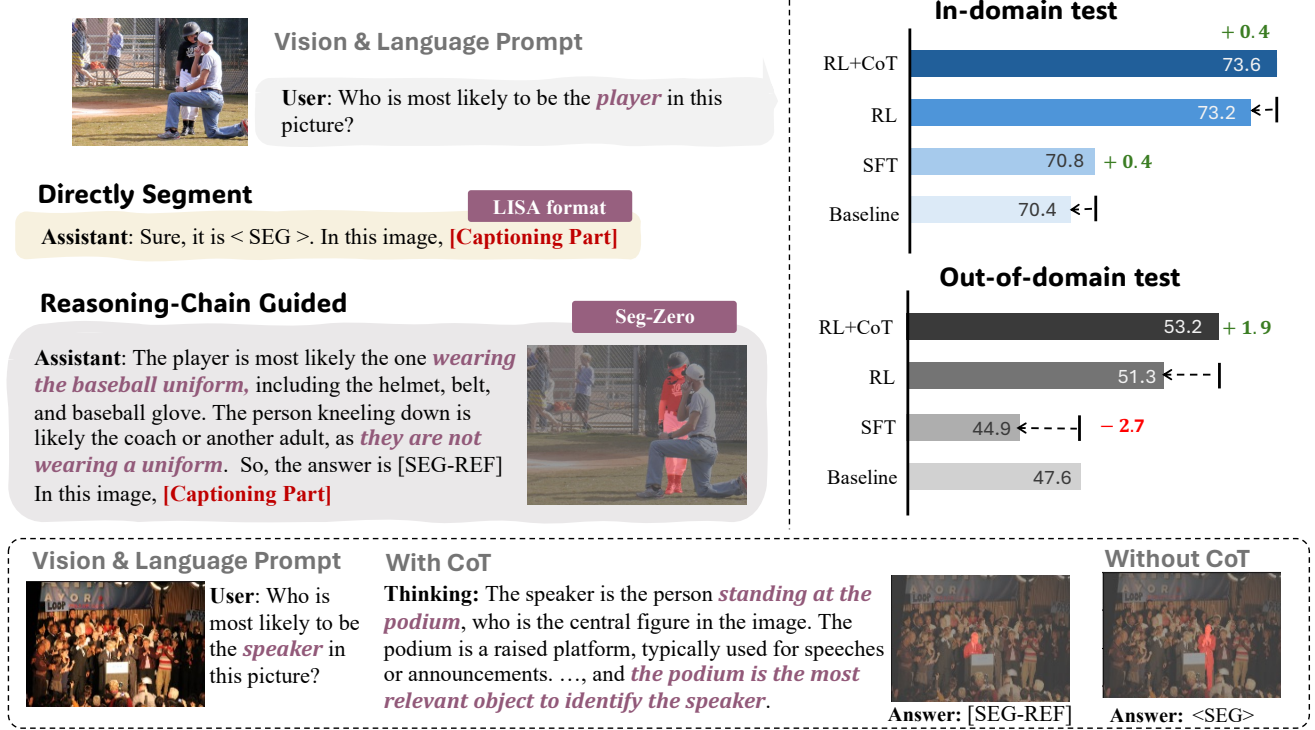https://github.com/dvlab-research/Seg-Zero

Figure 1. Seg-Zero generates a reasoning chain before producing the final segmentation mask. It utilizes a pure reinforcement learning (RL) strategy, learning the reasoning process from zero. In comparison to supervised fine-tuning (SFT), the RL-based model demonstrates superior performance on both in-domain and out-of-domain data, and the integration of reasoning chain further enhances its effectiveness.

Abstract

Traditional methods for reasoning segmentation rely on supervised fine-tuning with categorical labels and simple descriptions, limiting its out-of-domain generalization and lacking explicit reasoning processes. To address these limitations, we propose Seg-Zero, a novel framework that demonstrates remarkable generalizability and derives explicit chain-of-thought reasoning through cognitive reinforcement. Seg-Zero introduces a decoupled architecture consisting of a reasoning model and a segmentation model. The reasoning model interprets user intentions, generates explicit reasoning chains, and produces positional prompts, which are subsequently used by the segmentation model to generate precious pixel-level masks. We design a sophisticated reward mechanism that integrates both format and accuracy rewards to effectively guide optimization directions. Trained exclusively via reinforcement learning with GRPO and without explicit reasoning data, Seg-Zero achieves robust zero-shot generalization and exhibits emergent test-time reasoning capabilities. Experiments show that Seg-Zero-7B achieves a zero-shot performance of 57.5 on the ReasonSeg benchmark, surpassing the prior LISA-7B by 18%. This significant improvement highlights Seg-Zero's ability to generalize across domains while presenting an explicit reasoning process. Code is available at https://github.com/dvlab-research/Seg-Zero.

1. Introduction

Reasoning segmentation generates pixel-wise masks by interpreting implicit queries through logical reasoning. This task shows significant potential in real-world applications, such as robots. Unlike conventional segmentation tasks that rely on simple categorical labels (e.g., "person" or "car"), reasoning segmentation addresses more complex and nuanced queries, such as "identify food that provides sustained energy." Such queries require logical reasoning and the integration of cross-domain knowledge to produce accurate segmentation masks.

Early attempts [3, 17, 32], such as LISA [17], have explored the use of multimodal large language models (MLLMs) to enhance reasoning segmentation capabilities. These methods bridge the gap between MLLMs and segmentation models by leveraging implicit semantic tokens. However, typical methods [7, 17, 32] rely solely on supervised fine-tuning (SFT) applied to mixed datasets containing only simple categorical information or basic factual descriptions [12, 13, 43]. Although this paradigm effectively aligns MLLMs [23, 24, 40] with segmentation models [14] in specific datasets, we observe that it lacks generalization capabilities. This can be demonstrated by: (i) Although existing methods excel on in-domain data, their performance significantly degrades on out-of-distribution (OOD) samples. (ii) SFT inevitably leads to catastrophic forgetting of general capabilities. (iii) The lack of an explicit reasoning process hinders their effectiveness in complex scenarios. These limitations motivate us to enhance general segmentation capabilities and improve reasoning performance by integrating an explicit reasoning process.

Recent studies [11] demonstrate that training with pure reinforcement learning (RL) activates the emergent test-time reasoning process, highlighting that reward-driven optimization is effective in enhancing model reasoning ability. Moreover, this approach often promotes generalization rather than overfitting to specific datasets. Inspired by this, we introduce Seg-Zero, a novel framework designed to enhance reasoning and cognitive capabilities for reasoning segmentation. Seg-Zero adopts a decoupled architecture, including a reasoning model and a segmentation model. The reasoning model is an MLLM capable of processing both image and user instructions. It outputs not only region-level bounding boxes (bbox) but also pixel-level points to precisely localize the target object. Subsequently, the segmentation model utilizes the bbox and points to produce pixel-level segmentation masks.

During training, we employ pure reinforcement learning, specifically GRPO [34], to fine-tune the reasoning model while keeping the segmentation model frozen. Rather than constructing datasets with explicitly annotated reasoning processes, we investigate the self-evolution potential of MLLM to develop reasoning capabilities, thereby achieving emergent reasoning from zero. To achieve this, we develop a sophisticated reward mechanism to enhance the reasoning process and regulate the output. These reward functions comprise two types: format rewards, which enforce constraints on the structure of the reasoning process and segmentation outputs, and accuracy rewards, which are calculated based on intersection over union (IoU) and L1 distance metrics. As illustrated in Figure 1, by leveraging optimized reward-driven reinforcement learning, our Seg-Zero exhibits emergent test-time reasoning abilities, similar to those demonstrated in LLMs [11, 27]. This reasoning process enables the model to effectively handle complex instructions by breaking them down into sequential analytical steps, thus achieving the precise localization of target objects. Seg-Zero demonstrates exceptional performance on both in-domain and OOD data, significantly exceeding the model trained through SFT. Furthermore, Seg-Zero maintains robust visual QA capability, without the need for VQA training data.

Experimental results show that, with only 9,000 training samples derived from RefCOCOg [43], our Seg-Zero-7B exhibits strong test-time reasoning capabilities and achieves superior generalization performance compared to models of the same scale. It achieves a zero-shot performance of 57.5 on ReasonSeg [17], surpassing the previous LISA-7B by 18%.

We summarize our contributions as follows:

- We propose Seg-Zero, a novel architecture designed for reasoning segmentation. Through the pure RL algorithm, Seg-Zero exhibits emergent reasoning abilities.
- We present a detailed comparison between SFT and RL, as well as the integration of reasoning chain. Results demonstrates that RL, combined with the reasoning chain, consistently enhances model performance.
- Extensive experiments demonstrate the effectiveness of our design and offer valuable insight for fine-tuning models using RL.

2. Related Works

2.1. Reasoning in Large Models

In recent years, Large Language Models (LLMs) have exhibited remarkable reasoning capabilities. By extending the length of the Chain-of-Thought (CoT) reasoning process, OpenAI-o1 [27] introduces inference-time scaling, significantly enhancing its reasoning performance. In the research community, several studies have attempted to achieve test-time scaling through various approaches, including process-based reward models [20, 38, 39], reinforcement learning (RL) [15, 34], and search algorithms [10, 37]. Notably, the recent DeepSeek-R1 [11], which utilizes the GRPO [34] algorithm, achieves superior performance with only a few thousand RL training steps. Building on advancements in

the LLMs community, several recent works have attempted to leverage the reasoning capabilities of MLLMs [16, 36]. For example, Open-R1-Multimodal [16] emphasizes mathematical reasoning, while R1-V [36] demonstrates exceptional performance in counting tasks. However, these works primarily address high-level reasoning and do not consider fine-grained pixel-level understanding of images. To fill this gap, our Seg-Zero is designed to enhance pixel-level reasoning through reinforcement learning.

2.2. Semantic Segmentation with Reasoning

Semantic segmentation aims to predict segmentation masks for specific classes. Numerous studies [1, 4, 5, 8, 21, 25, 33, 44], including DeepLab [6], MaskFormer [9] and SAM [14] have made significant progress in this task, making it a well-addressed problem. Instead of segmenting objects with explicit class labels, referring expression segmentation [13, 43] focuses on segmenting target objects based on short, explicit text queries. This task is more challenging as images often contain multiple instances of the same object class with varying attributes, requiring the model to identify and segment the instance that most closely matches the text query. LISA [17] advances this field further by introducing the reasoning segmentation task. In this task, the text queries are either more intricate or significantly longer, demanding models with strong reasoning capabilities to accurately interpret and segment the target objects.

2.3. MLLMs for Segmentation

Since LISA [17, 41] introduced the ' <SEG >' token to bridge the gap between MLLMs and segmentation models, several subsequent works [3, 7, 32] have explored the use of MLLMs for segmentation tasks. Most of these approaches, including OneTokenSegAll [3] and PixelLM [32], follow LISA's paradigm by utilizing special tokens to connect MLLMs with segmentation models. However, this design necessitates extensive data to fine-tune both the MLLM and the segmentation decoder, and may even compromise the pixel precious of the original segmentation models. Our proposed Seg-Zero also employs a decoupled design for ease of adoption, while further leveraging the reasoning ability of MLLMs to achieve superior results.

3. Method

In this section, we introduce our Seg-Zero model and the associated reinforcement learning framework. We first describe how we address the segmentation problem in Section 3.1. Next, we present the architecture of the Seg-Zero in Section 3.2. Finally, we describe the reward functions (Section 3.3) and training details (Section 3.4) in the reinforcement learning framework.

Figure 2. Illustration of our RL training process. In this case, the model generates three samples by itself, calculates the rewards, and optimizes towards samples that achieve higher rewards.

3.1. Pipeline Formulation

Given an image $\mathbf{I}$ and a label $\mathbf{T}$, the segmentation task aims to produce a binary segmentation mask $\mathbf{M}$ that accurately identifies the region corresponding to $\mathbf{T}$. The label $\mathbf{T}$ can vary in complexity, ranging from a simple class label (e.g., "bird"), to a straightforward phrase (e.g., "woman in blue"), or even to long and intricate expressions (e.g., "The unusual thing in the image"). The latter two types of expressions require the model to perform reasoning to accurately segment the most relevant objects.

Inspired by recent advancements in the reasoning capabilities of large models [11, 34, 36], we leverage this ability to develop a pipeline for reasoning-based segmentation. Specifically, we decouple the reasoning process and the segmentation process. We first employ reinforcement learning to a MLLM to activate its reasoning ability, enabling it to generate reasoning process and produce accurate bounding box $\mathbf{B}$ and two points $\mathbf{P_1}, \mathbf{P_2}$ that best localize the target object. These bounding box and points are then used as prompts for SOTA segmentation models [14, 30] to produce fine-grained segmentation masks. Seg-Zero is trained using
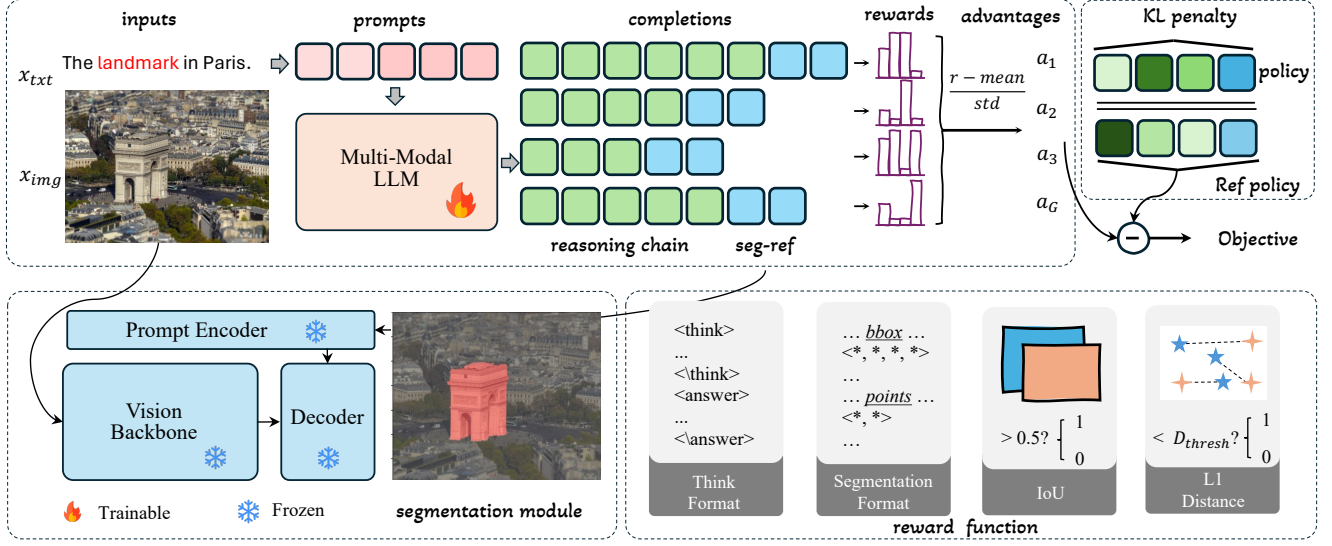
Figure 3. Seg-Zero includes a reasoning model and a segmentation model. The reasoning model is a MLLM that generates a reasoning chain and provides segmentation prompts. Subsequently the segmentation model produces pixel-wise mask.

reinforcement learning, as illustrated in Figure 2.

3.2. Seg-Zero Model

Current MLLMs [2, 18, 24, 40, 45] exhibit impressive performance in processing multi-modal inputs but are unable to generate fine-grained segmentation masks. Conversely, modern segmentation models [14, 30] provides fine-grained segmentation ability but lack robust reasoning capabilities. To bridge this gap, we propose Seg-Zero, a framework that includes a reasoning model and a segmentation model. Additionally, we introduce the novel strategy to effectively activate the reasoning ability of MLLM within the framework. Its whole architecture is shown in Figure 3.

Reasoning Model. We employ Qwen2.5-VL [2] as our reasoning model $\mathcal{F}_{reason}$. While Qwen2.5-VL demonstrates exceptional performance in object detection by predicting the bbox, this region-level bbox is insufficient for providing more fine-grained pixel-level localization. Unlike object detection, segmentation requires a more precise understanding of pixel-level details, as multiple objects may exist within a single bounding box. Therefore, in addition to bounding box, we also incorporate points that lie within the target object to improve localization accuracy. During the reinforcement learning stage, the format rewards are employed to ensure the model generates structured outputs, which are subsequently processed by a post-processing function $\mathcal{G}$ to extract the bounding box $\mathbf{B}$ and the two points $\mathbf{P_1}, \mathbf{P_2}$. This process can be formulated as:

$$\mathbf{B}, \mathbf{P_1}, \mathbf{P_2} = \mathcal{G}(\mathcal{F}_{reason}(\mathbf{I}, \mathbf{T})). \quad (1)$$

Segmentation Model. Modern segmentation models

[14, 30] accept various types of prompts, including bounding boxes and points, to generate accurate segmentation masks. We employ SAM2 [30] as our segmentation model $\mathcal{F}_{seg}$ due to its superior performance and efficient inference speed. Leveraging the bounding boxes and points provided by the reasoning model, the segmentation model can generate precise, fine-grained mask for the target object. This process can be formally expressed as:

$$\mathbf{M} = \mathcal{F}_{seg}(\mathbf{B}, \mathbf{P_1}, \mathbf{P_2}). \quad (2)$$

Test-time Reasoning. Reasoning is the crucial part in reasoning segmentation tasks. Inspired by DeepSeek-R1-Zero, we intentionally avoid using any explicit Chain-of-Thought (CoT) data to teach Seg-Zero reasoning skills. Instead, we aim to activate its reasoning capabilities from zero, enabling the model to autonomously generate a logical CoT before producing the final answer. To achieve this, we design a structured user prompt and a sophisticated reward mechanism that guides the reasoning model to follow specific instructions. As shown in Figure 4, this user prompt instructs Seg-Zero to analyze and compare objects in the image, beginning by generating a reasoning process, followed by the final answer in a predefined format.

3.3. Reward Functions

Reward functions play a pivotal role in reinforcement learning, as they determine the optimization directions of the model. We manually design the following five reward functions for reinforcement learning.

Thinking Format Reward. This reward is designed to force the model engage in a structured thinking process.

4

Figure 4. User prompt for Seg-Zero. '{Question}' is replaced with object description $\mathbf{T}$ in the training and inference.

It guides the model output its reasoning steps within the <think> and </think>tags, and the final answer is included between the <answer> and </answer>tags.

Segmentation Format Reward. Different from counting or other QA tasks, segmentation task is highly dependent on the format of the answer. We provide two types of segmentation format rewards: soft and strict. Under soft constrains, if the keywords *bbox* and *points* appear in the answer, and their corresponding values consist of four and two coordinates, respectively, the format is considered correct. Under strict constraints, the format is only considered correct if the model outputs exact keywords (e.g., *bbox*, *points_1*, *points_2*) in the required structure.

Bbox IoU Reward. This reward evaluates the IoU between the predicted bbox against the ground-truth bbox. A reward of 1 is assigned if their IoU greater than 0.5; otherwise, the reward is 0.

Bbox L1 Reward. This reward evaluates the L1 distance between the predicted bbox and the ground-truth bbox. A reward of 1 is assigned if their L1 distance less than 10 pixels; otherwise, the reward is 0.

Point L1 Reward. This reward evaluates the L1 distance between predicted points and the ground-truth points. We first determine whether the predicted points are inside the bounding box. Then the reward is set to 1 if the minimal distance between the predicted points and the ground-truth points is less than 100 pixels; otherwise, the reward is 0.

3.4. Training

We build the training data from public available segmentation datasets and train our Seg-Zero using GRPO algorithm.

Data Preparation. The training data is generated using the original mask annotations from existing referring expression segmentation datasets (e.g., RefCOCOg [43]). Based on the mask, we extract the leftmost, topmost, rightmost, and bottommost pixels of the mask to generate the bounding box $\mathbf{B}$. Additionally, we compute the center points of the two largest inscribe circles within the mask, denoted as $\mathbf{P_1}$ and $\mathbf{P_2}$. Consequently, the ground truth data comprises the bbox coordinates $[\mathbf{B}_{x1}, \mathbf{B}_{y1}, \mathbf{B}_{x2}, \mathbf{B}_{y2}]$ and the coordinates of the two center points $[\mathbf{P_1}_x, \mathbf{P_1}_y]$ and

Model	Type	CoT	RefCOCOg	ReasonSeg
Baseline			70.4	47.6
Seg-Zero	SFT	×	70.8	44.9
Seg-Zero	RL	×	73.2	51.3
Seg-Zero	RL	✓	**73.6**	**53.8**

Table 1. Segmentation task comparison. Model trained with RL + CoT thinking reward achieves best performance on in-domain and OOD data.
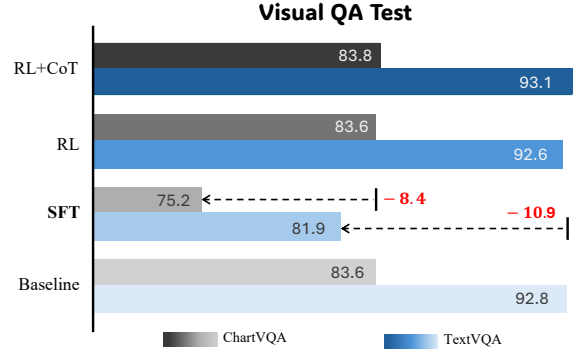

Figure 5. Visual QA task comparison. SFT suffers catastrophic forgetting, while RL preserves general Visual QA ability.

$[\mathbf{P_2}_x, \mathbf{P_2}_y]$. We do not incorporate any CoT processing into the training data. To ensure consistency across different datasets, all images are rescaled to a uniform resolution of 840x840 pixels.

GRPO. We do not include any *reasoning data* for a cold-start training process to teach the model's reasoning ability. Instead, we let our Seg-Zero learn from zero. Specifically, we initiate training directly from the pre-trained Qwen2.5-VL-3B model, utilizing the aforementioned rewards and applying the GRPO algorithm [34]. We illustrate oue RL training process in Figure 2.

Figure 6. Qualitative Results on ReasonSeg [17]. The reason chain helps analyze user instructions and segment the correct objects.

4. Experiment

4.1. Experimental Settings

Datasets. We training our Seg-Zero with only 9,000 samples adopted from RefCOCOg, using data preparation strategy mentioned in Section 3.4. The test data includes RefCOCO(+/g) [43] and ReasonSeg [17].

Implementation Details. We employ Qwen2.5-VL-3B [2] and SAM2-Large [30] as our default reasoning model and segmentation model, respectively. Seg-Zero is trained on an 8xH200 GPU server using the DeepSpeed [29] library. During training, we use a total batch size of 16 with a sampling number of 8 per training step. The initial learning rate is set to 1e-6 and the weight decay is 0.01.

Evaluation Metrics. Following previous works [13, 43], we calculate gIoU and cIoU. The gIoU is the average of all per-image Intersection-over-Unions (IoUs), while the cIoU calculates the cumulative intersection over the cumulative union. Unless specified, we use gIoU as our default metric, as it equally considers both large and small objects.

4.2. SFT vs. RL

We compare the performance of SFT and RL. The baseline model is Qwen2.5-VL-3B + SAM2. For the non-CoT setting, we eliminate the thinking format reward, thus the model does not generate a CoT reasoning process before outputting the final answer. Our comparison includes both in-domain and OOD segmentation tasks [26, 35], as well as general QA tasks. The corresponding results are shown in Table 1, Figure 1 and Figure 5.

SFT vs. RL without CoT. From the first two rows in Table 1, we observe that on the in-domain dataset RefCOCOg, SFT achieves nearly the same performance as the baseline model. This may be because of the strong baseline performance of the original Qwen2.5-VL-3B. However, its performance significantly declines on the OOD ReasonSeg dataset, suggesting that SFT negatively impacts the model's generalization ability. In contrast, comparing the first and third rows, we find that RL consistently improves performance on both in-domain and OOD datasets, demonstrating the effectiveness of RL. Besides, from Figure 5, we observe that the SFT model suffers from catastrophic forgetting of its original visual QA ability, while the RL model effec-

Model	Bbox	Points	RefCOCOg	ReasonSeg
Baseline			70.4	47.6
Seg-Zero	✗	✓	69.6	45.5
Seg-Zero	✓	✗	72.9	53.6
Seg-Zero	✓	✓	**73.6**	**53.8**

Table 2. Ablation on the design of bbox and points prompt.

Model	Type	RefCOCOg	ReasonSeg	sum
Seg-Zero	Soft	70.2	**54.1**	124.3
Seg-Zero	Hard	**73.6**	53.8	**127.4**

Table 3. Ablation on the accuracy reward type.

Model	Type	RefCOCOg	ReasonSeg	sum
Seg-Zero	Soft	**73.6**	53.8	127.4
Seg-Zero	Strict	73.0	**56.1**	**129.1**

Table 4. Ablation on the format reward type. Strict format is better.

Reasoning Model	RefCOCOg	ReasonSeg
Qwen2-VL-2B	70.1	37.2
Qwen2.5-VL-3B	73.0	56.1
Qwen2.5-VL-7B	**74.2**	**57.5**

Table 5. Ablation on reasoning model choice. Larger scale model achieves better performance.

Figure 7. Changes in completion length during training. Larger scale model tends to generate longer response.

tively preserves this capability.

RL without CoT vs. RL with CoT. From the last two rows in Table 1, we find that both RL and RL with CoT achieves superior performance on both the in-domain RefCOCOg and OOD ReasonSeg datasets, significantly outperforming the baseline. This indicates that RL effectively boost the models' capabilities. However, with CoT, our Seg-Zero demonstrates even better performance compared to its counterparts without CoT, indicating the reasoning process enhances the models ability to handle OOD data samples. From Figure 5, it is noteworthy that the introduction of CoT reasoning leads to a slight performance improvement in visual QA tasks for models trained with RL.

4.3. Ablation Study

We conduct several ablation studies to verify the effectiveness of our design. For the ablation study, the default settings are as follows: we perform reinforcement learning using the GRPO algorithm on 9,000 samples and evaluate the model on the RefCOCOg test and the ReasonSeg test.

Design of Bbox and Points. Table 2 demonstrates the effectiveness of our bbox and points prompt design. We observe that using only point prompts results in worst performance. When both bbox and point prompts are utilized, Seg-Zero achieves its best performance, indicating that the combination of these prompts enhances pixel-level localization accuracy.

Soft vs. Hard Accurary Rewards. In Section 3.3, we describe the bbox IoU reward, the bbox L1 reward, and the point L1 reward. We apply specific thresholds to convert these metrics into binary rewards. Additionally, we conduct ablation studies on soft counterparts. For the bbox IoU reward, we directly use the IoU value as the soft reward. For the L1-based rewards, we define the soft reward as $1 - \frac{L1\_dist}{max\{image\_size\}}$. From Table 3, we observe that while the soft reward achieves a minor improvement in ReasonSeg, it significantly underperforms compared to the hard

reward on RefCOCOg.

Soft vs. Strict Format Rewards. In Section 3.3, we introduce two types of segmentation format reward: the soft and strict. From Table 4, we find that the strict format reward significantly improves performance gain on OOD data in ReasonSeg. Through qualitative analysis of the training steps, we find that the strict format reward progresses slowly in the initial stages, as it is more challenging to sample formats that precisely match the strict criteria. However, as training step increases, model with strict format reward tend to output longer response.

Reasoning Model Scale. We conduct an ablation study on reasoning models of varying scales, ranging from 2B to 7B parameters, under the same rewards and training settings. As shown in Table 5, we observe that model perfor-

Figure 8. Changes in rewards during training. We show the mean value across a batch.

| Method | ReasonSeg | | | |
| | val | | test | |
	gIoU	cIoU	gIoU	cIoU
OVSeg	28.5	18.6	26.1	20.8
ReLA	22.4	19.9	21.3	22.0
Grounded-SAM	26.0	14.5	21.3	16.4
LISA-7B-LLaVA1.5	53.6	52.3	48.7	48.8
LISA-13B-LLaVA1.5	57.7	60.3	53.8	50.8
SAM4MLLM	46.7	48.1	-	-
Qwen2.5VL-3B+SAM2	53.8	44.1	47.6	37.4
Seg-Zero-3B (ours)	62.6	58.5	56.1	48.6
Seg-Zero-7B (ours)	**62.6**	**62.0**	**57.5**	**52.0**

Table 6. Zero-shot reasoning segmentation results.

| Method | refCOCO | refCOCO+ | refCOCOg |
	testA	testA	test
LAVT	75.8	68.4	62.1
ReLA	76.5	71.0	66.0
LISA-7B	76.5	67.4	68.5
PixelLM-7B	76.5	71.7	70.5
PerceptionGPT-7B	78.6	73.9	71.7
Seg-Zero-3B (ours)	79.3	73.7	71.5
Seg-Zero-7B (ours)	**80.3**	**76.2**	**72.6**

Table 7. Referring expression segmentation results. We compare the cIoU in this table.

longer responses. As the training progresses, the minimal completion length gradually increases. However, there is a drop in average completion length during the initial few steps. By analyzing the output during the training process, we find that this occurs because the model initially prioritizes learning the correct output format, which often results in shorter responses. Once the format reward saturates, the model shifts its focus toward generating answers with higher accuracy, leading to longer and more detailed responses.

We visualize the changes in rewards during the training process. As shown in Figure 8, the format rewards converge to 1 in few steps, and the accuracy rewards gradually increase over time. This suggests that the format rewards initially dominates the optimization direction, leading to a decrease in response length during the initial training steps, as shown in Figure 7. However, as the format rewards converge and the accuracy rewards gradually increases, the model's completion length (i.e., the CoT reasoning process) begins to expand.

4.4. Comparison with Other Methods

In this part, we train our Seg-Zero using hard accuracy rewards and strict format rewards. The sampling number is set to 16. And we only train our Seg-Zero on

mance on both in-domain and OOD data improves as the model scale increases.

Changes in Completion Length. Figure 7 illustrates the trends in completion lengths across different model sizes. The results indicate that larger model tends to generate

9,000 samples from RefCOCOg. We compare OVSeg [19], Grounded-SAM [31], LISA [17], SAM4MLLM [7], LAVT [42], ReLA [22], PixelLM [32], PerceptionGPT [28].

Reasoning Segmentation. We compare the zero-shot performance on ReasonSeg [17], results are shown in Table 6. We can find our Seg-Zero achieves the SOTA zero-shot performance acrosss various methods.

Referring Expression Segmentation. The results on referring expression segmentation is shown on Table 7. Moreover, we find that the ground-truth annotations in Ref-COCO(+/g) are not precise enough, which suggests that our Seg-Zero model should, in principle, achieve better performance than values in the table. *Supplementary materials* provide detailed analysis.

4.5. Qualitative Results

We provide several examples in Figure 6 and Figure 9. We can easily obeserve that the reasoning process is helpful in analyzing user instructions, especially when there are multiple objects within the same class categories. For instance, Seg-Zero demonstrates its ability to discern that the a 'recreational vehicle' is more appropriate than a 'truck' in the context of a 'road trip', and correctly indentifies that a 'conductor' is 'positioned at the front of the stage'.

5. Conclusion

In this paper, we propose Seg-Zero, an novel framework that integrates CoT reasoning process into segmentation tasks. We design a sophisticated reward mechanism, incorporating both format and accuracy constraints, to guide the optimization directions. By training exclusively with RL, Seg-Zero emerges reasoning capabilities without relying on any supervised reasoning data. We present a detailed comparison between SFT and RL, as well as the introduction of reason chain. Additionally, we offer insightful perspectives on the design of RL and the reward functions.
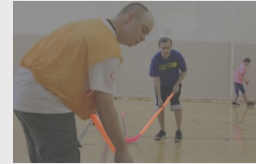
User

In ice hockey, players use specific equipment to *control the puck and score goals*. What object in the picture could be used by the players for this purpose?
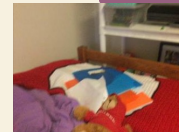
Seg-Zero

Thinking : The object in question is a *hockey stick*, which is used by players to control the puck and score goals in ice hockey. The hockey stick is a *long, thin, and flexible stick* that players use to hit the puck. The hockey stick is typically made of *wood or composite materials* and has a handle and a blade. The blade is used to control the puck and the stick is used to hit the puck.
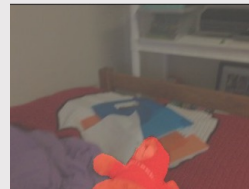
User

Children often like to sleep with *their favorite toys* by their side. What object in the picture can be served as a child's favorite toy?

Seg-Zero

Thinking : The object in question is a child's favorite toy, which is typically *a stuffed animal or a soft toy*. Among the objects visible in the image, *the teddy bear* fits this description.

User

During leisure time, what object in the living room can *create an immersive audio-visual experience* when paired with other electronic objects?

Seg-Zero

Thinking : The question asks for an object that can create an immersive audio-visual experience when paired with other electronic objects. The speaker, which is a device used to *amplify sound*, fits this description. It is a common object found in *living rooms and can be paired with other electronic devices* like a television or a sound system to create an immersive audio-visual experience.
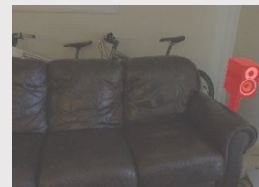
User

Something that protects the knees.

Seg-Zero

Thinking : The object that protects the knees is a *knee pad*, which is a piece of protective gear typically worn on the knees to prevent injuries during physical activities. The knee pad in the image *is black and fits snugly around the knee area*.

Figure 9. More examples.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 6
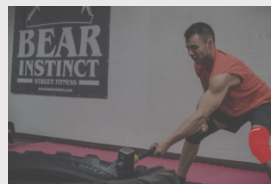
[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2025. 2, 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[7] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multimodal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2, 3, 9

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Perpixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 3

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[10] Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*, 2023. 2

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3

[12] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, highquality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 3, 6

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3, 4

[15] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024. 2

[16] EvolvingLMMs Lab. Open R1 Multimodal. https://github.com/EvolvingLMMs-Lab/open-r1-multimodal, 2025. 3

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 3, 6, 9

[18] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 4

[19] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 9

[20] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for highresolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3

[22] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. 9

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 4

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015. 3

[26] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 6

[27] OpenAI. OpenAI o1. https://openai.com/o1/, 2024. 2

[28] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024. 9

[29] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020. 6

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4, 6

[31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 9

[32] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 2, 3, 9

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3, 5

[35] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[36] R1-V Team. R1-V. https://github.com/Deep-Agent/R1-V?tab=readme-ov-file, 2025. 3

[37] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. 2

[38] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. 2

[39] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023. 2

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 4

[41] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 3

[42] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022. 9

[43] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 3, 5, 6

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[45] Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, et al. Lyra: An efficient and speech-centric framework for omni-cognition. *arXiv preprint arXiv:2412.09501*, 2024. 4