# Reasoning via Implicit Self-supervised Emergence for Instruction Segmentation

**Qing Zhou**[1], **Lichang Yang**[1], **Yuyu Jia**[1], **Junyu Gao**[1], **Weiping Ni**[2*], **Junzheng Wu**[2], **Qi Wang**[1*]

[1]School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, China
[2]Department of Remote Sensing, Northwest Institute of Nuclear Technology, China
{mrazhou, ylc, jyy2019, gjy3035}@mail.nwpu.edu.cn, {niweiping, wujunzheng}@nint.ac.cn, crabwq@gmail.com

## Abstract

We challenge the assumption that complex instruction-guided segmentation tasks necessitate equally complex and explicit supervision. This paper introduces RISE (**R**easoning via **I**mplicit **S**elf-supervised **E**mergence), a framework that learns intricate compositional reasoning, spanning spatial relations to world knowledge, without a single ground-truth mask. To achieve this, RISE employs reinforcement learning with GRPO guided by a single, strikingly simple reward: the semantic alignment score between the textual instruction and the predicted image region. Our primary discovery is the implicit emergence of a high-quality chain-of-thought process from this minimalist signal. Within a structured format, the model autonomously learns to understand instructions by accessing its latent knowledge, inferring spatial relationships—capabilities inherent in its architecture but unlocked by our simple objective. Remarkably, our emergent reasoning yields highly competitive results: RISE achieves 58.7 gIoU on the ReasonSeg benchmark, on par with methods using geometric rewards. Furthermore, we show extreme data efficiency: a variant trained on only 2,000 ImageNet-label pairs establishes a new state-of-the-art for annotation-free referring segmentation with 79.6 cIoU on RefCOCO.

## 1 Introduction

Instruction-guided segmentation aims to produce pixel-perfect masks for objects described by natural language commands, encompassing both simple referring expressions (Yu et al. 2016; Kazemzadeh et al. 2014) and complex reasoning-based queries (Lai et al. 2024; Yang et al. 2023). While early work successfully tackled straightforward instructions like *"the red car"*, the frontier is rapidly advancing towards compositional instructions that demand a deeper understanding, such as *"someone who is most likely to be the player"* or *"someone with bare arms"*. Answering such queries requires more than mere object recognition; it necessitates a model's ability to perform intricate compositional reasoning, blending visual perception, spatial awareness, and even world knowledge.

The dominant paradigms for this challenge, illustrated in Figure 1, have been supervised learning. Seminal works like

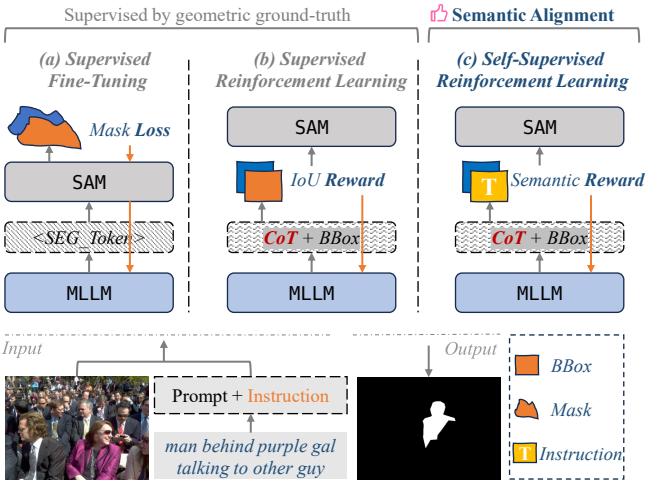


Figure 1: A comparison of learning paradigms for instruction-guided segmentation. **(a) Supervised Fine-Tuning** (e.g., LISA) is bottlenecked by its direct reliance on costly ground-truth masks for computing a *Mask Loss*. **(b) Supervised RL** (e.g., Seg-Zero) still depends on geometric ground-truth to calculate an *IoU Reward*, failing to break the annotation dependency. **(c) Our Self-Supervised RL (RISE)** introduces a new paradigm, using a *Semantic Reward* derived from the original instruction (**T**).

LISA (Lai et al. 2024) pioneered the use of Multimodal Large Language Models (MLLMs) (Shang et al. 2024; Bai et al. 2025a; Li 2024) fine-tuned on large-scale datasets with paired instruction-mask annotations (Fig. 1a). While effective, this supervised fine-tuning (SFT) approach suffers from high annotation costs and often struggles with out-of-domain generalization (Liu et al. 2025a). To improve reasoning capabilities, more recent methods, such as Seg-Zero (Liu et al. 2025a), have turned to reinforcement learning (RL) to generate a chain-of-thought (CoT) (Wei et al. 2022; Ma et al. 2025) (Fig. 1b). However, a critical limitation persists: **their reward functions remain fundamentally anchored to ground-truth geometric annotations**, such as bounding box IoU or point-based distances. This reliance on explicit geometric supervision, fails to break the core dependency on costly, human-provided geometric data.

*Co-corresponding authors

This reliance on geometric supervision begs a fundamental question: *Can complex, compositional reasoning emerge from a purely semantic learning signal, entirely devoid of explicit geometric guidance?* In this work, we provide a definitive and affirmative answer. We introduce **RISE** (**R**easoning via **I**mplicit **S**elf-supervised **E**mergence), a new self-supervised RL paradigm (Fig. 1c). RISE employs reinforcement learning with GRPO (Shao et al. 2024) but revolutionizes the learning signal. It completely obviates the need for any geometric annotations, guided by a single, strikingly simple reward: **the semantic alignment score between the instruction and the predicted image region.**

Our central and most striking discovery is that a sophisticated, compositional reasoning process implicitly emerges, driven solely by a *non-geometric, semantic reward* and without any supervision on the *reasoning process itself*. To consistently maximize the semantic similarity score for a query, the model autonomously discovers the effective strategy is to perform instruction-relevant reasoning and reasoning-relevant grounding. This involves accessing its latent knowledge to ground abstract concepts, resolving spatial relationships between objects, and synthesizing these intermediate findings to identify the final target. This entire chain-of-thought is not designed into the reward function or taught via examples; it is an emergent solution that the model converges upon as the effective path to maximizing a simple, holistic semantic objective.

Through this emergent process, RISE achieves remarkable performance, demonstrating results that are on par with its fully supervised counterparts on challenging segmentation benchmarks. Our contributions are threefold:

- We propose RISE, a novel framework that learns to perform intricate compositional reasoning for instruction segmentation without any mask-based supervision.
- We demonstrate, the emergence of a chain-of-thought reasoning process from a simple, holistic text-image semantic alignment objective, revealing a powerful way to harness the latent capabilities of large models.
- We show our annotation-free approach is highly effective, achieving competitive performance against supervised methods and establishing a new state-of-the-art among annotation-free methods, even when trained on minimal classification-level data.

## 2 Related Work

### Reasoning in Vision-Language Models

The capabilities of MLLMs have expanded dramatically, enabling complex interactions between vision and text (Lai et al. 2024; Shang et al. 2024; Huang, Zhang, and Li 2025; Bai et al. 2025a). A key catalyst for unlocking their advanced cognitive abilities has been the CoT prompting technique (Ma et al. 2025; Wei et al. 2022), which elicits step-by-step reasoning. Building on this, recent research has focused on teaching models to generate reasoning without explicit CoT supervision in the training data. A prominent approach is to use RL to reward desirable outcomes, thereby incentivizing the model to produce a coherent reasoning process as an intermediate step. For instance, works like DeepSeek-R1 (Guo et al. 2025) have shown that RL, particularly with algorithms like GRPO, can effectively cultivate emergent reasoning capabilities in LLMs on text-based tasks. Our work, RISE, builds upon this RL-for-reasoning paradigm. However, while prior works focus on language-only tasks or use RL with geometrically-grounded rewards for visual tasks (Liu et al. 2025a), we demonstrate that a purely semantic, spatially-agnostic reward is sufficient to elicit compositional reasoning for instruction-guided segmentation.

### Instruction-Guided Segmentation

Instruction-guided segmentation aims to ground natural language commands in pixel space. The task has evolved from referring expression segmentation, which focuses on localizing objects based on relatively simple descriptive phrases (Yu et al. 2016; Kazemzadeh et al. 2014), to the more challenging reasoning segmentation. Pioneered by LISA (Lai et al. 2024), reasoning segmentation involves interpreting complex instructions that require multi-step logic, world knowledge, and an understanding of relationships between multiple objects. Subsequent works have improved upon LISA's architecture and training schemes (Ren et al. 2024b; Chen et al. 2024; Bai et al. 2025b), but they overwhelmingly share a common foundation: a reliance on SFT using datasets with meticulously annotated pixel-level masks. This dependency on dense supervision creates a significant scalability bottleneck, as highlighted in our introduction. Recent efforts like Seg-Zero (Liu et al. 2025a) have moved towards RL to reduce the dependency on explicit CoT annotations, but critically, their reward mechanism still requires geometric ground truth (e.g., IoU scores) for optimization. In contrast, RISE fundamentally breaks this dependency, learning to segment from instructions without any form of geometric or mask-based supervision.

### Learning from Annotation-Free Supervision

The prohibitive cost of dense pixel-wise annotations has long motivated research into weakly-supervised and self-supervised methods. In semantic segmentation, early efforts explored learning from weaker signals like image-level labels (Zhou et al. 2016; Ahn and Kwak 2018), bounding boxes (Khoreva et al. 2017) or few-shot labels (Jia et al. 2025a,b). With the advent of powerful vision-language models like CLIP (Radford et al. 2021), a new frontier of self-supervised visual grounding has emerged. These methods leverage the semantic alignment between image patches and text descriptions to perform tasks like open-vocabulary object detection or localization without explicit bounding box annotations (Liu et al. 2024). However, these approaches are typically limited to grounding simple nouns or short phrases and do not address the complex, compositional instructions that require multi-step reasoning. They excel at answering "*what*", but struggle with "*why*" or "*how*" as embedded in a complex command. RISE bridges this crucial gap. RISE demonstrate that the simple, annotation-free signal of text-image semantic similarity, when combined with an RL framework, is powerful enough to unlock not just simple object grounding, but the emergence of sophisticated

compositional reasoning required for challenging segmentation tasks.

## 3 Preliminaries

To clearly situate our contribution, we first formalize the three dominant learning paradigms for instruction-guided segmentation. Let $I$ be an input image, $T$ be a textual instruction, and $\theta$ represent the model parameters.

### Supervised Fine-Tuning

The SFT paradigm, exemplified by methods like LISA (Lai et al. 2024), treats the task as a direct, supervised mapping problem. A model $f_\theta$ is trained to predict a segmentation mask $M_{pred}$ given the image-instruction pair. The optimization objective is to minimize a pixel-level loss function against a ground-truth mask $M_{gt}$:

$$\min_\theta \mathcal{L}_{mask}(f_\theta(I, T), M_{gt}) \tag{1}$$

where $\mathcal{L}_{mask}$ can be a Dice or Binary Cross-Entropy loss. The core of this paradigm is its reliance on direct, dense, and costly pixel-level supervision.

### Supervised Reinforcement Learning

To elicit more complex reasoning, recent works like Seg-Zero (Liu et al. 2025a) formulate the task using RL. The model acts as a policy $\pi_\theta$ that, given a state $s = (I, T)$, generates an action $a$, which typically includes a CoT text and a predicted bounding box, $a = (\text{CoT}, B_{pred})$. The policy is optimized to maximize the expected total reward:

$$\max_\theta J(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[R_g(a)] \tag{2}$$

Crucially, the reward function $R_g$ in this paradigm is fundamentally anchored to geometric ground-truth. For example, it is often based on the IoU between the predicted box $B_{pred}$ and the ground-truth box $B_{gt}$:

$$R_g(a) = \text{IoU}(B_{pred}, B_{gt}) \tag{3}$$

Despite removing the need for explicit CoT supervision, it fails to break the dependency on geometric annotations.

### Self-Supervised Reinforcement Learning (Ours)

We propose a new paradigm that retains the powerful framework of RL but revolutionizes the reward signal to achieve self-supervision. The optimization objective remains the maximization of expected reward as in Eq. 2. However, we introduce a self-supervised reward function $R_s$ that is completely independent of any geometric ground-truth. Instead, it is derived solely from the semantic consistency between the input itself:

$$R_s(a) = \text{Sim}(I_{crop}(B_{pred}), T) \tag{4}$$

where $\text{Sim}(\cdot, \cdot)$ is a semantic similarity function and $I_{crop}(B_{pred})$ is the image region defined by the predicted box. This paradigm breaks the reliance on geometric annotations, and detailed in the following section.

## 4 Method: RISE

As shown in Figure 2, RISE operates within a RL framework designed to elicit compositional reasoning from a self-supervised signal. We detail its architecture, problem formulation, and learning mechanism below.

### Overall Architecture

The architecture of RISE is composed of two main components: a reasoning agent and a segmentation executor.

- **Reasoning Agent:** We employ a MLLM (Bai et al. 2025a) as the core reasoning agent. Its role is to process the input image $I$ and instruction $T$ and to generate a textual output that includes an explicit CoT and a set of localization proposals. These proposals consist of a primary bounding box ($B_{pred}$) and two auxiliary interior points ($P_1, P_2$). Only this MLLM is trainable and is optimized via RL.
- **Segmentation Executor:** We use a pre-trained, frozen Segment Anything Model (SAM) (Ravi et al. 2024) as our segmentation executor. It takes the full set of localization proposals ($B_{pred}, P_1, P_2$) generated by the MLLM as prompts to produce the final, fine-grained segmentation mask $M_{pred}$. By keeping SAM frozen, we isolate the learning process entirely within the reasoning agent.

### Problem Formulation

We formally define the instruction segmentation task within a standard RL framework:

- **State ($s$):** A tuple of the input image and the textual instruction, $s = (I, T)$.
- **Action ($a$):** The full text sequence generated by the MLLM. The action is structured to contain a thinking process (CoT) and a final answer with localization proposals, $a = (\text{CoT}, B_{pred}, P_1, P_2)$.
- **Policy ($\pi_\theta$):** The MLLM itself, parameterized by $\theta$, which defines a probability distribution over possible actions given a state, $\pi_\theta(a|s)$.
- **Reward ($R$):** A scalar feedback signal that evaluates the quality of the action $a$. This reward is the core of our self-supervised approach, as detailed next.

### Self-supervised Reward Mechanism

Our reward mechanism is designed to be completely free of geometric annotations, relying instead on semantic consistency and structural correctness.

**Semantic Reward ($R_{sem}$)** The primary learning signal, $R_{sem}$, measures the semantic alignment between the content of the predicted region and the original instruction, based on the insight that a correct localization must contain visual concepts that match the text. The process is as follows: 1. The bounding box $B_{pred}$ is parsed from the action $a$. 2. This box is used to crop the corresponding region $I_{crop}$ from the input image $I$. 3. A pre-trained and frozen vision-language Model, such as CLIP (Radford et al. 2021), is used to compute the cosine similarity between the embeddings of the

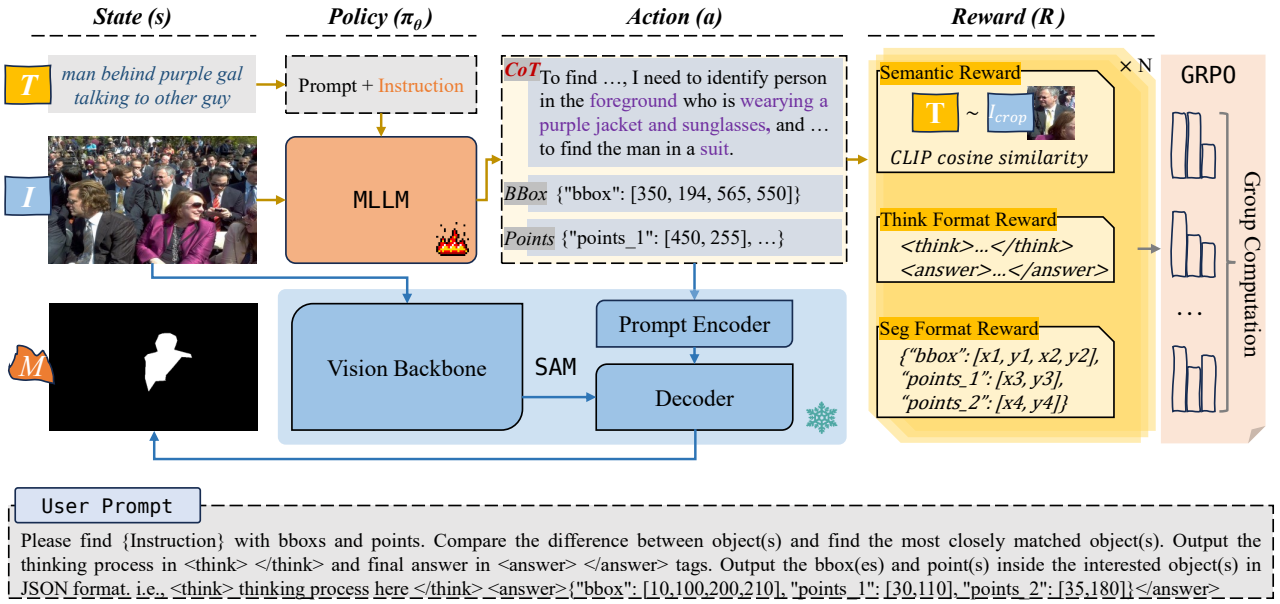


Figure 2: An overview of the RISE framework, illustrating its distinct training and inference stages. During the **training stage (top and right paths)**, given an image-instruction pair as the state $s$, the MLLM policy $\pi_\theta$ generates an action $a$ containing a CoT and localization proposals (a bounding box and points). This action is then evaluated by our self-supervised reward function, which consists of a primary Semantic Reward and auxiliary Format Rewards. The resulting scalar reward is used by the GRPO algorithm to update the MLLM policy. Notably, this entire training loop is self-contained and does not involve the segmentation model. During the **inference stage (bottom path)**, the trained MLLM generates its proposals, which are then fed into a frozen SAM to produce the final, fine-grained segmentation mask $M$.

cropped image and the instruction text.

$$R_{sem} = \frac{\text{emb}_I(I_{crop}) \cdot \text{emb}_T(T)}{||\text{emb}_I(I_{crop})|| \cdot ||\text{emb}_T(T)||} \quad (5)$$

This reward mechanism is designed to be inherently spatially-agnostic. The similarity function (e.g., CLIP) evaluates only the content of the isolated patch $I_{crop}$, which is decontextualized from its original position in the full image. Consequently, the reward signal provides feedback on *what* is inside the proposed box but offers no direct gradient or information regarding *where* the box should be located or how its position should be corrected. This design choice forces the learning process to rely solely on discerning the semantic content that best matches the full instruction. To achieve a high reward for a complex, multi-component instruction, the model is therefore incentivized to generate a bounding box that precisely captures the unique entity satisfying all aforementioned textual descriptions, as this is the most direct way to maximize the semantic alignment score.

**Format Rewards ($R_{format}$)** To ensure the MLLM generates syntactically valid and usable actions, we incorporate two binary format rewards.

- **Thinking Format Reward ($R_{think}$):** A reward of 1 is assigned if the generated text correctly includes the <think>...</think> and <answer>...</answer> structures; otherwise, the reward is 0.
- **Segmentation Format Reward ($R_{seg}$):** A reward of 1 is assigned if the action's answer contains a parsable bounding box and two points in the correct JSON format with valid coordinate values. While the geometric accuracy of these points is not rewarded, compelling the model to generate them has been empirically found to improve the quality of the primary bounding box proposal, likely by encouraging a more detailed internal representation of the target object's geometry.

**Total Reward** The final reward used for policy optimization is the direct sum of the components:

$$R_{total} = R_{sem} + R_{think} + R_{seg} \quad (6)$$

The learning process is primarily driven by the self-supervised semantic reward $R_{sem}$, while the format rewards provide a structured scaffold for the model's output.

## Policy Optimization with GRPO

To optimize the policy $\pi_\theta$, we employ the GRPO(Shao et al. 2024) algorithm. As shown in Figure 2, the training process for each step involves sampling a set of N completions (actions) from the current policy $\pi_\theta$. For each completion, we compute the total reward $R_{total}$. GRPO then uses the ranked preferences of these (action, reward) pairs to construct a reward model implicitly and update the policy. This process effectively guides the MLLM to assign higher probabilities to action sequences that yield higher rewards, thereby encouraging the discovery and refinement of the emergent reasoning strategies that lead to maximal semantic alignment.

| Method | ReasonSeg val gIoU | ReasonSeg val cIoU | ReasonSeg test gIoU | ReasonSeg test cIoU |
|---|---|---|---|---|
| ReLA (Liu, Ding, and Jiang 2023) | 22.4 | 19.9 | 21.3 | 22.0 |
| Grounded-SAM (Ren et al. 2024a) | 26.0 | 14.5 | 21.3 | 16.4 |
| LISA-7B-LLaVA1.5 (Lai et al. 2024) | 53.6 | 52.3 | 48.7 | 48.8 |
| LISA-13B-LLaVA1.5 (Lai et al. 2024) | 57.7 | 60.3 | 53.8 | 50.8 |
| SAM4MLLM (Chen et al. 2024) | 46.7 | 48.1 | - | - |
| Qwen2.5VL-3B + SAM2 | 53.8 | 44.1 | 47.6 | 37.4 |
| Qwen2.5VL-7B + SAM2 | 57.6 | 48.3 | 50.1 | 41.2 |
| Seg-Zero-3B$^{\dagger}$ (Liu et al. 2025b) | 62.6 | 58.5 | 56.1 | 48.6 |
| Seg-Zero-7B$^{\dagger}$ (Liu et al. 2025b) | 62.6 | 62.0 | 57.5 | 52.0 |
| Seg-Zero-3B$^{*}$ (Liu et al. 2025b) | 59.1 | 48.8 | 52.5 | 43.4 |
| Seg-Zero-7B$^{*}$ (Liu et al. 2025b) | 61.6 | 52.6 | 58.2 | 52.4 |
| **RISE-3B-ClsOnly (ours 💡)** | 53.5 | 38.6 | 50.4 | 41.7 |
| **RISE-7B-ClsOnly (ours 💡)** | 57.1 | 47.8 | 54.5 | 44.0 |
| **RISE-3B (ours)** | 52.6 | 43.1 | 52.9 | 45.1 |
| **RISE-7B (ours 🏆)** | **62.0** | **55.3** | **58.7** | **52.5** |

Table 1: Comparison with SOTA methods on the **Zero-shot ReasonSeg**. $^{\dagger}$ denotes scores reported in the original paper, while $^{*}$ denotes reproduction. 🏆 denotes the best method. 💡 indicates a variant of RISE trained only with 2K classification data (ImageNet).

| Method | RCO testA | RCO+ testA | RCOg test |
|---|---|---|---|
| *Supervised* | | | |
| LAVT (Yang et al. 2022) | 75.8 | 68.4 | 62.1 |
| ReLA (Liu, Ding, and Jiang 2023) | 76.5 | 71.0 | 66.0 |
| LISA-7B (Lai et al. 2024) | 76.5 | 67.4 | 68.5 |
| PixelLM-7B (Ren et al. 2024b) | 76.5 | 71.7 | 70.5 |
| MagNet (Chng et al. 2024) | 78.3 | 73.6 | 69.3 |
| PerceptionGPT-7B (Pi et al. 2024) | 78.6 | 73.9 | 71.7 |
| Seg-Zero-3B$^{\dagger}$ (Liu et al. 2025b) | 79.3 | 73.7 | 71.5 |
| Seg-Zero-7B$^{\dagger}$ (Liu et al. 2025b) | 80.3 | 76.2 | 72.6 |
| Seg-Zero-3B$^{*}$ (Liu et al. 2025b) | 76.0 | 70.6 | 68.8 |
| Seg-Zero-7B$^{*}$ (Liu et al. 2025b) | **79.4** | **73.7** | **73.2** |
| *Self-supervised* | | | |
| **RISE-3B-ClsOnly (ours 💡)** | 76.3 | 71.8 | 69.4 |
| **RISE-7B-ClsOnly (ours 💡)** | 79.6 | 77.6 | 73.1 |
| **RISE-3B (ours 🧪)** | 76.5 | 72.1 | 68.9 |
| **RISE-7B (ours 🧪)** | **79.7** | **77.7** | **73.4** |

Table 2: Performance on **Referring Expression Segmentation** (cIoU %). 🧪 denotes performance on par with some supervised methods.

# 5 Experiments

## Experimental Setup

**Datasets.** We evaluate RISE across a comprehensive suite of benchmarks to assess its reasoning capabilities. For Reasoning Segmentation, we use the challenging ReasonSeg benchmark (Lai et al. 2024), which contains complex, multi-step instructions. For Referring Expression Segmentation, we use the standard RefCOCO(+/g) (Yu et al. 2016) datasets.

**Implementation Details.** RISE employs Qwen2.5VL-3B as the Reasoning Agent and SAM2-Large as the Segmentation Executor. The semantic reward $R_{sem}$ is computed using the ViT-B/32 CLIP model. For training optimization, we adopt the GRPO algorithm with 8 sampled completions per instance, learning rate 1e-6, maintaining consistency with the Seg-Zero configuration. For testing on RefCOCO (+/g), use the official prompt consistent with that of Qwen2.5VL.

**Training Data for RISE.** RISE is trained on a small dataset of just 2,000 image-instruction pairs sampled from the RefCOCOg training split. This data contains only simple referring expressions, not complex reasoning instructions. A single checkpoint trained on this limited data is used for evaluation across all benchmarks, using only the image-text pairs without their geometric annotations.

**Training Data for RISE-ClsOnly.** To demonstrate the extreme data efficiency of our paradigm, our RISE-ClsOnly variant is trained on a remarkably minimal dataset: 2,000 random images from ImageNet-1K (Russakovsky et al. 2015), where the instruction for each image is simply its **class label**. This simulates a realistic, low-cost scenario of learning from only widely available classification data.

## Main Results

**Performance on Reasoning Segmentation** We first evaluate RISE on the challenging ReasonSeg benchmark, which requires complex compositional reasoning. As shown in Table 1, self-supervised RISE-7B achieves a gIoU of 58.7 on the test set. This result is highly competitive, performing on par with the fully supervised reinforcement learning method, Seg-Zero-7B$^{*}$ (58.2 gIoU), which relies on ground-truth geometric rewards. This is a central finding of our work: by leveraging a purely semantic, spatially-agnostic reward, RISE can elicit a level of complex reasoning comparable to methods that use explicit geometric supervision. This finding suggests that the rich, latent reasoning structures within pre-trained MLLMs can be effectively activated and steered for complex visual tasks using only a semantic signal, without resorting to direct geometric supervision. Furthermore, classification-only variant, RISE-7B-ClsOnly, also shows strong performance, validating the robustness.

**Performance on Referring Expression Segmentation** We next evaluate RISE on the standard referring expression segmentation benchmarks to assess its performance on more traditional referring tasks.

Table 2 compares RISE with fully supervised methods. While not surpassing the state-of-the-art, our self-supervised RISE-7B achieves a cIoU of 73.4 on RefCOCOg, demonstrating strong and competitive performance against several supervised methods. This indicates that the reasoning capabilities unlocked by our method also translate effectively to simpler object referring tasks.

When compared against other annotation-free methods in Table 3, RISE establishes a new state-of-the-art. The performance of our RISE-ClsOnly models is particularly compelling, demonstrating data efficiency. For instance, on RefCOCO, the 3B variant achieves a cIoU of **76.3**, which is

| Method | RCO testA | RCO+ testA | RCOg test |
|---|---|---|---|
| *Zero-shot Methods* | | | |
| BSAP (Wang et al. 2024) | 27.0 | 27.8 | 34.5 |
| SAM-CLIP (Ni et al. 2023) | 25.8 | 28.0 | 38.9 |
| Ref-Diff (Ni et al. 2023) | 38.4 | 40.5 | 44.5 |
| TAS (Suo, Zhu, and Yang 2023) | 41.1 | 49.1 | 46.8 |
| CaR (Sun et al. 2024) | 35.4 | 36.0 | 36.6 |
| IteRPrimE (Wang et al. 2025) | 46.5 | 51.6 | 45.8 |
| *Large VLMs* | | | |
| Qwen2.5VL-3B + SAM2 | 75.9 | 71.5 | 70.1 |
| Qwen2.5VL-7B + SAM2 | 77.8 | 73.5 | 71.2 |
| **RISE-3B-ClsOnly (ours 💡🏆🚀)** | 76.3 | 71.8 | 69.4 |
| **RISE-7B-ClsOnly (ours 💡🏆🚀)** | **79.6** | **77.6** | **73.1** |

Table 3: Performance on Zero-shot **Referring Expression Segmentation**. RISE significantly outperforms all previous self-supervised and weakly-supervised approaches by a large margin (🚀), establishing a new state-of-the-art for annotation-free visual grounding.

| Method | $R_{sem}$ | $R_{think}$ | $R_{seg}$ | gIoU | cIoU |
|---|---|---|---|---|---|
| Format-only | | ✓ | ✓ | 28.6 | 21.1 |
| Semantic-only | ✓ | | | 51.5 | 44.0 |
| RISE (w/o Seg Format) | ✓ | ✓ | | 50.2 | 42.5 |
| RISE (w/o Think Format) | ✓ | | ✓ | 48.7 | 41.1 |
| **RISE** (Full) | ✓ | ✓ | ✓ | **52.9** | **45.1** |

Table 4: Ablation study on the components of our self-supervised reward function on ReasonSeg testset. The results show that the semantic reward ($R_{sem}$) is the primary driver of performance.

| Method | CoT | gIoU | cIoU |
|---|---|---|---|
| RISE-ClsOnly (w/o CoT) | | 41.3 | 32.0 |
| RISE-ClsOnly | ✓ | **50.4** | **41.7** |
| RISE (w/o CoT) | | 47.6 | 38.5 |
| RISE | ✓ | **52.9** | **45.1** |

Table 5: Ablation study on the impact of Chain-of-Thought.

already slightly better than the strong Qwen2.5VL-3B baseline (75.9), and the 7B variant achieves **79.6**. While trained on only 2,000 generic and out-of-domain ImageNet-label pairs, this variant demonstrates significantly superior performance. This highlights the effectiveness and robustness of our self-supervised RL paradigm in learning powerful visual grounding capabilities from minimal, low-cost data.

## Ablation Studies and Analysis

**Analysis of Reward Components** We first investigate the contribution of each component in our self-supervised reward function. The results are presented in Table 4. The "Format-only" setting, which lacks the semantic reward ($R_{sem}$), performs poorly, confirming that $R_{sem}$ **is the primary driver of learning**. The "Semantic-only" setting performs reasonably well, but is outperformed by the full model, indicating that the format rewards provide a crucial structural scaffold. Removing either the think format reward ($R_{think}$) or the segmentation format reward ($R_{seg}$) leads to a drop in performance, particularly on the more complex ReasonSeg benchmark. Specifically, removing $R_{think}$ causes a drop from 45.1 cIoU (Full RISE) to 41.1 cIoU, and removing $R_{seg}$ leads to 42.5 cIoU. This validates our design of a composite reward where a dense semantic signal is complemented by sparse structural guidance.

**The Importance of Chain-of-Thought** Next, we analyze the impact of explicitly generating a Chain-of-Thought. In Table 5, we compare our full model against a variant where the model is not prompted to produce a <think>...</think> block. The results show a dramatic performance drop on ReasonSeg (from 45.1 to 38.5 for the full model) when CoT is removed. Similarly, the RISE-ClsOnly variant drops from 41.7 to 32.0. This confirms that the emergent reasoning process, when externalized as an explicit CoT, is a critical mechanism for the model to deconstruct and solve complex, multi-step instructions. The structured thinking process is not merely an artifact, but a functional component for successful task completion.

**Auxiliary Points.** We also study the effect of generating auxiliary points alongside the primary bounding box. The results are shown in Table 6. Interestingly, removing the points leads to a noticeable drop on the more complex ReasonSeg (from 45.1 to 42.8). Recall that these points receive no geometric reward; they are only required for structural format correctness. This finding suggests an intriguing hypothesis: compelling the model to generate these additional points, which often correspond to salient parts of the target object, may encourage a more detailed and fine-grained internal geometric representation within the MLLM. This refined internal understanding proves particularly beneficial for complex reasoning tasks, even though the points themselves are not directly supervised for accuracy.

## Qualitative Analysis

To provide a more intuitive understanding of how our self-supervised framework fosters emergent reasoning, we present a qualitative comparison in Figure 3. We compare the RISE model against a baseline variant trained only with format rewards ($R_{think}$ and $R_{seg}$), which lacks the crucial semantic guidance from $R_{sem}$.

The top example showcases a task requiring both attribute and spatial reasoning. Given the instruction *"someone with bare arms,"* the baseline model defaults to a simplistic heuristic, incorrectly grounding the most salient person in the scene and making a flawed inference (*"...likely the one sitting on the motorcycle, as they are the only one with no visible clothing on their arms"*). In stark contrast, RISE generates a sophisticated and correct reasoning chain. It correctly infers that *"bare arms"* is visually associated with a *"sleeveless shirt"* and accurately uses the spatial cue *"left motorcycle"* to disambiguate between subjects, leading to the correct final segmentation.

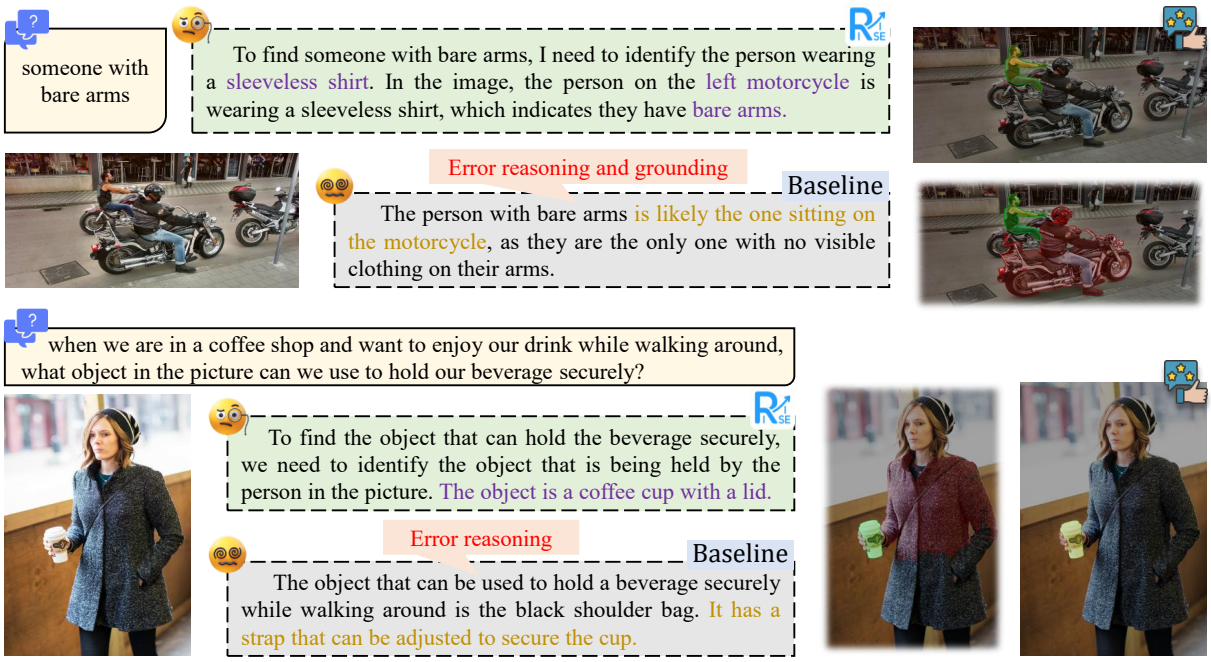


Figure 3: Qualitative comparison of the reasoning processes between RISE and a baseline. The baseline here refers to a variant of our model trained without the semantic reward ($R_{sem}$), relying only on format rewards. **Top Example (Spatial & Attribute Reasoning):** For the instruction *"someone with bare arms,"* the baseline incorrectly focuses on the most prominent person and makes a flawed assumption. In contrast, RISE correctly identifies the person wearing a *"sleeveless shirt"* on the *"left motorcycle"* as the target, demonstrating accurate attribute and spatial reasoning. **Bottom Example (Affordance & World Knowledge Reasoning):** Given a functional query about holding a beverage, the baseline exhibits error reasoning, incorrectly identifying a "black shoulder bag" based on a superficial feature ("strap"). RISE, however, correctly reasons about the object's function and affordance, identifying the "coffee cup with a lid" as the logical answer.

| Method | Points | gIoU | cIoU |
|---|---|---|---|
| RISE (w/o Points) | | 52.2 | 42.8 |
| RISE | ✓ | **52.9** | **45.1** |

Table 6: Ablation study on the impact of Points.

The bottom example highlights reasoning about object affordance and world knowledge. When asked for an object to *"hold our beverage securely,"* the baseline exhibits classic error reasoning. It fixates on a superficial feature—a *"strap"* on a *"shoulder bag"*—and incorrectly associates it with the function of *"securing a cup."* This demonstrates a failure to grasp the instruction's true intent. RISE, however, correctly reasons about the functional requirements. Its CoT reveals a logical process of identifying what the person is holding and recognizing it as a *"coffee cup with a lid,"* which is the only object suitable for the requested function.

These case studies provide strong, tangible evidence of the reasoning capabilities unlocked by our training paradigm. The baseline, lacking semantic guidance during training, often converges to simplistic, error-prone heuristics. In contrast, RISE, having been optimized to maximize a holistic semantic alignment score, demonstrates a consistently more robust and human-like reasoning process at inference time. This suggests that our training objective successfully creates a pressure for the model to develop a deeper, compositional understanding of language and its visual grounding, as this constitutes a more effective strategy to succeed during its self-supervised training.

## 6 Conclusion

In this work, we introduced RISE, a self-supervised RL framework that demonstrates a fundamental principle: sophisticated, multi-step reasoning can emerge from a simple, spatially-agnostic semantic objective, without any geometric supervision. By rewarding only the holistic alignment between an instruction and a visual region, our model is intrinsically motivated to develop an internal chain-of-thought, proving to be a highly effective strategy. This emergent reasoning allows RISE to achieve competitive performance on complex reasoning tasks and establish a new state-of-the-art for annotation-free referring segmentation. Our findings suggest a paradigm shift towards harnessing the latent capabilities of large pre-trained models through simple, fundamental objectives, rather than relying on complex, explicit supervision. Future work could explore extending this paradigm to multi-object scenarios and other visual tasks.

## Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 62471394 and U21B2041, 62306241, U62576284.

## References

Ahn, J.; and Kwak, S. 2018. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In *CVPR*, 4981–4990.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bai, Z.; He, T.; Mei, H.; Wang, P.; Gao, Z.; Chen, J.; Zhang, Z.; and Shou, M. Z. 2025b. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 37: 6833–6859.

Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024. SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In *ECCV*, 323–340. Springer.

Chng, Y. X.; Zheng, H.; Han, Y.; Qiu, X.; and Huang, G. 2024. Mask grounding for referring image segmentation. In *CVPR*, 26573–26583.

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

Huang, S.; Zhang, H.; and Li, X. 2025. Enhance Vision-Language Alignment with Noise. 17449–17457. Association for the Advancement of Artificial Intelligence.

Jia, Y.; Fu, W.; Gao, J.; and Wang, Q. 2025a. Dual-View Classifier Evolution for Generalized Remote Sensing Few-Shot Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–10.

Jia, Y.; Zhou, Q.; Gao, J.; and Wang, Q. 2025b. Entity-Guided Attention Twisting Network for Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–10.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 787–798.

Khoreva, A.; Benenson, R.; Hosang, J. H.; Hein, M.; and Schiele, B. 2017. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *CVPR*, 1665–1674.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*, 9579–9589.

Li, X. 2024. Positive-Incentive Noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 8708–8714.

Liu, C.; Ding, H.; and Jiang, X. 2023. GRES: Generalized Referring Expression Segmentation. In *CVPR*, 23592–23601.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *ECCV*, volume 15105, 38–55.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025a. Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement. *arXiv preprint arXiv:2503.06520*.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025b. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Ma, X.; Wan, G.; Yu, R.; Fang, G.; and Wang, X. 2025. CoT-Valve: Length-Compressible Chain-of-Thought Tuning. *arXiv preprint arXiv:2502.09601*.

Ni, M.; Zhang, Y.; Feng, K.; Li, X.; Guo, Y.; and Zuo, W. 2023. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*.

Pi, R.; Yao, L.; Gao, J.; Zhang, J.; and Zhang, T. 2024. Perceptiongpt: Effectively fusing visual perception into LLM. In *CVPR*, 27124–27133.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024a. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

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

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. In *CVPR*, volume 115, 211–252.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

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

Sun, S.; Li, R.; Torr, P.; Gu, X.; and Li, S. 2024. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, 13171–13182.

Suo, Y.; Zhu, L.; and Yang, Y. 2023. Text augmented spatial-aware zero-shot referring image segmentation. *EMNLP*.

Wang, H.; Zhan, Y.; Liu, L.; Ding, L.; Yang, Y.; and Yu, J. 2024. Towards Alleviating Text-to-Image Retrieval Hallucination for CLIP in Zero-shot Learning. *arXiv preprint arXiv:2402.18400*.

Wang, Y.; Ni, J.; Liu, Y.; Yuan, C.; and Tang, Y. 2025. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In *AAAI*, volume 39, 8159–8168.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837.

Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2023. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 18155–18165.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *ECCV*, 69–85. Springer.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*, 2921–2929.