

# Skip the Talk, Re-Focus on Vision: Latent Reasoning for Reasoning Segmentation in Multimodal Large Language Models

Tianhang Guo<sup>1,\*</sup> Yulin He<sup>1,\*</sup> Wei Chen<sup>1,†</sup>

Wenjuan Zhou<sup>1</sup> Yuhang Li<sup>1</sup> Xinbiao Gan<sup>1</sup>

<sup>1</sup>National University of Defense Technology, Changsha, China

\*Equal contribution. †Corresponding author.

Email: guotianhang.\_@nudt.edu.cn

## Abstract

Reasoning segmentation aims to interpret implicit textual queries and enable fine-grained visual perception, which is critical for applications such as human-computer interaction and embodied agents. Existing methods typically generate explicit Chain-of-Thought (CoT) by multimodal large language models (MLLMs) before localizing the target. Although intuitive, such explicit verbal reasoning introduces substantial attention interference: redundant textual tokens disrupt attention during perception-token generation and also increase the effective distance between visual tokens. To address this issue, we propose LIRSeg, which fully replaces explicit CoT with a compact set of learnable latent tokens for reasoning segmentation. LIRSeg is trained in two stages: spatial alignment grounds the latent tokens in object-relevant visual evidence, and GRPO further optimizes them with segmentation rewards. To make these compact latent tokens more informative, we introduce three complementary mechanisms from an information perspective: extreme-advantage sampling for selecting informative training signals, decoupled exploration-stability updates for learning complementary representations, and latent diversity amplification for preventing representational collapse. Extensive experiments on benchmarks demonstrate that LIRSeg consistently improves both segmentation accuracy and reasoning efficiency. Compared with the VisionReasoner baseline, LIRSeg achieves absolute gIoU improvements of 4.9% on ReasonSeg, 7.1% on MUSE, and 4.7% on MMR, while achieving a  $\sim 16\times$  reduction in reasoning tokens. Code is available in supplementary materials.

## Introduction

Multimodal large language models (MLLMs) (Liu et al. 2023; Bai et al. 2025) unify language generation with visual perception, achieving performance on open-world image-text understanding. Beyond recognizing visual content, reasoning segmentation (Lai et al. 2024) demands a perception capability from MLLMs: given an image and a free-form query, models need to infer the intent and localize the target at the pixel level. Unlike referring segmentation, where the target is explicitly named, queries in reasoning segmentation are implicit and require the model to integrate functional, contextual, and intentional cues to identify objects.

Existing methods broadly fall into two paradigms. Supervised fine-tuning (SFT) methods, represented by LISA (Lai et al. 2024), combine an MLLM with a segmentation model

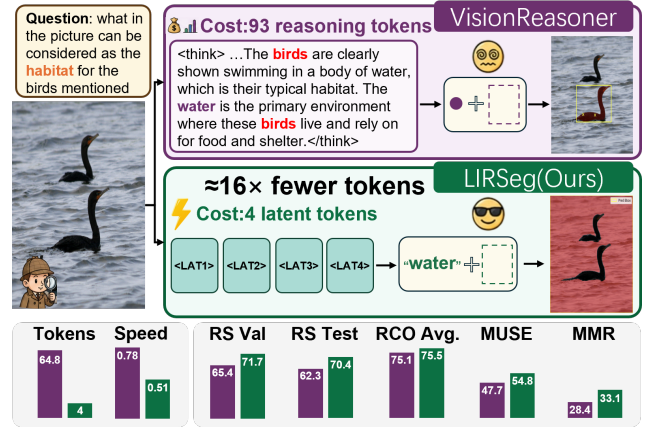


Figure 1: **Motivation.** LIRSeg implicitly reasons with few latent tokens, cutting reasoning tokens by  $\sim 16\times$  while improving accuracy on ReasonSeg, RefCOCO, MUSE, and MMR, unlike explicit-CoT methods that distract from targets.

and train on annotated data. While effective in-distribution, they often generalize poorly to out-of-distribution (OOD) due to limited reasoning traces in annotated data. Reinforcement learning (RL) methods, such as Seg-Zero (Liu et al. 2025a), instead optimize the MLLM with Group Relative Policy Optimization (GRPO) (Shao et al. 2024) under perception rewards, enabling self-generated reasoning and improved OOD generalization. However, RL-based methods are prone to overthinking due to uncontrolled CoT traces (Wang et al. 2025). To curb this overthinking problem, PixelThink (Wang et al. 2025) uses an extra large-scale expert MLLM to estimate query difficulty and cap reasoning length. DR<sup>2</sup>Seg (He et al. 2026) decomposes the task into two passes and rewards a short, self-contained description. These methods shorten the reasoning chain, yet leave its fundamental form unchanged: reasoning is still explicitly verbalized in natural language, allowing discrete text tokens and distractor texts to interfere with attention during perception-token generation.

A natural idea is therefore to remove verbalized intermediate reasoning entirely. However, prior studies (Liu et al. 2025a,b) have shown that simply removing the reasoning process can degrade model generalization. We therefore ex-

plore the potential of latent reasoning (Hao et al. 2024) in reasoning segmentation. Beyond general latent reasoning in LLMs, latent reasoning segmentation requires latent tokens to jointly encode reasoning semantics and visual-spatial information, while serving as a bridge to generate pixel-level masks, making the task substantially more challenging.

To this end, we propose LIRSeg (Latent Implicit Reasoning for Segmentation), a fully latent reasoning framework for reasoning segmentation. LIRSeg is trained in two stages. In the first stage, SFT aligns learnable latent tokens with object-relevant visual regions, grounding latent reasoning in pixel-level evidence. In the second stage, RL refines these latent tokens using segmentation rewards, enabling latent reasoning within a compact token budget. Unlike explicit CoT, which inherits rich representational diversity from pretrained MLLMs, latent reasoning with limited data lacks such inherent diversity. We address this limitation with three complementary information-theoretic mechanisms: extreme-advantage sampling to amplify informative learning signals, decoupled exploration–stability updates to preserve disentangled token representations, and latent diversity amplification to prevent representational collapse. As shown in Figure 1, LIRSeg achieves efficient reasoning and accurate perception without relying on thinking supervision.

Our contributions are summarized as:

- We propose LIRSeg, the first RL-optimized latent reasoning framework for reasoning segmentation, replacing redundant verbal CoT with learnable latent tokens that refocus MLLMs on task-relevant visual evidence.
- We propose three complementary mechanisms from an information-theoretic perspective: extreme-advantage sampling, decoupled exploration–stability updates, and latent diversity amplification, which jointly enhance the diversity and representational capacity of latent tokens.
- We validate the effectiveness of our method across diverse benchmarks and multiple model configurations pairing MLLMs with segmentation models, establishing a new state-of-the-art in efficient reasoning segmentation.

## Related Work

### Reasoning Segmentation

Referring expression segmentation localizes image regions from direct linguistic descriptions (Yang et al. 2022; Liu, Ding, and Jiang 2023a; Ding et al. 2021). Reasoning segmentation extends this setting to implicit queries, where the target need be inferred from contextual cues or semantic relations, requiring both reasoning and visual grounding.

Early approaches for reasoning segmentation primarily rely on SFT, pioneered by LISA (Lai et al. 2024) and extended along three directions: enhanced token designs and decoders (Ren et al. 2024; Bao et al. 2024; Yang et al. 2024; Qian et al. 2026), multi-round interactive segmentation (Wu et al. 2023; Wang et al. 2024; Lu et al. 2025), and unified visual instruction frameworks (Zhang et al. 2024; Wang and Ke 2024; Wei et al. 2024). However, their reasoning ability is largely inherited from annotated supervision, limiting generalization to unseen scenarios (Liu et al. 2025a).

More recently, RL has emerged as an alternative, pioneered by Seg-Zero (Liu et al. 2025a) and extended by follow-up works (Liu et al. 2025b; Huang et al. 2025; Zhu et al. 2025b; You and Wu 2025; Zhao et al. 2026). However, these methods still generate explicit chain-of-thought (CoT) before producing the mask, incurring substantial overhead and attention interference. Although PixelThink (Wang et al. 2025) and DR<sup>2</sup>Seg (He et al. 2026) adopt adaptive strategies, they still rely on explicit textual CoT, leaving latent reasoning for reasoning segmentation largely unexplored. This work demonstrates the feasibility of this paradigm and broadens its methodological scope.

## Latent Reasoning

Latent reasoning has been explored in LLMs through additional computation positions (Goyal et al. 2024), internal rationales (Zelikman et al. 2024), and continuous-state feedback (Hao et al. 2024; Deng, Choi, and Shieber 2024). Recent reasoning MLLMs further show that continuous representations can effectively preserve perceptual and reasoning information that is difficult to express in discrete language (Pham and Ngo 2025; Yang et al. 2025; Li et al. 2025).

Prior work provides both methodological and theoretical foundations for exploring this paradigm in reasoning segmentation. This work further investigates the unique challenges of latent reasoning in this setting. Unlike general MLLM reasoning, reasoning segmentation requires latent tokens to jointly encode semantic reasoning and visual-spatial information, while also serving as an interface for generating pixel-level masks. This tighter coupling between reasoning and dense prediction makes optimization substantially more challenging. These considerations motivate the design of LIRSeg, which achieves both accurate and efficient segmentation across diverse benchmarks and different combinations of MLLM backbones and segmentation models.

## Method

### Preliminaries on Reasoning Segmentation

Reasoning segmentation takes an image  $I$  and an implicit textual query  $Q$  as input and predicts a binary mask  $M$  for the target. Unlike referring segmentation,  $Q$  typically does not name the target explicitly, requiring the model to infer the intended object before localization. Existing methods typically follow an explicit chain-of-thought (CoT) paradigm. Given  $(I, Q)$ , the MLLM is queried and decodes a language reasoning sequence  $\mathcal{T} = (t_1, \dots, t_L)$  to analyze visual and contextual cues, and then generates structured output  $\mathcal{F}_{cot}$  in the format: `<think> $\mathcal{T}$ </think><answer> $\mathcal{A}$ </answer>` where  $\mathcal{T}$  denotes the explicit CoT and  $\mathcal{A}$  denotes the perception answer. This prompt is subsequently fed into the segmentation model  $S$  to produce the mask prediction  $M$ .

### LIRSeg: A Latent Reasoning Framework for Reasoning Segmentation

As shown in Figure 2, we extend the vocabulary of the MLLM with  $K$  new special tokens, denoted by  $\langle \text{LAT}_1 \rangle, \dots, \langle \text{LAT}_K \rangle$ . Their token embeddings are learnable and are initialized before training. The decoding prompt

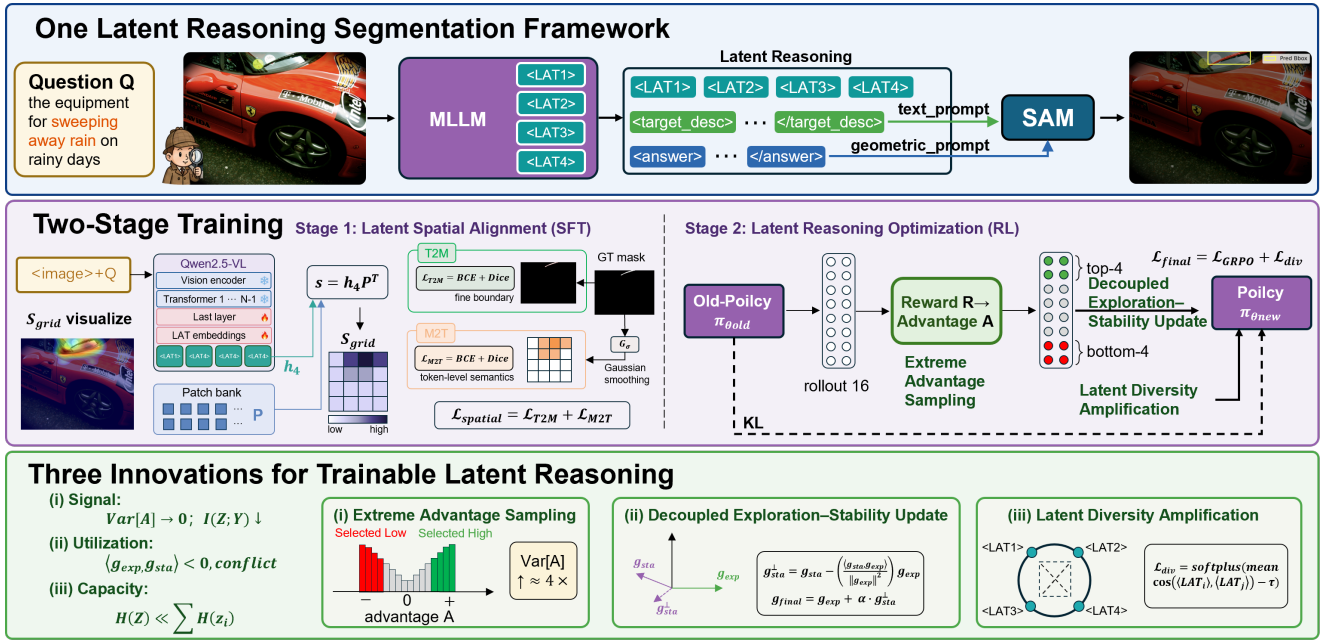


Figure 2: **Overview of LIRSeg.** *Top:* Latent reasoning via latent tokens produces a text description and bounding-box. *Middle:* Two-stage training: spatial alignment followed by RL optimization. *Bottom:* Three innovation components.

specifies a fixed output template that requires the model to generate these tokens consecutively at the beginning of its response, before producing the target description  $\mathcal{D}$  and the perception answer  $\mathcal{A}$ :  $\mathcal{F}_{lat} = \langle \text{LAT}_1 \rangle \cdots \langle \text{LAT}_K \rangle \langle \text{target\_desc} \rangle \mathcal{D} \langle \text{target\_desc} \rangle \langle \text{answer} \rangle \mathcal{A} \langle \text{answer} \rangle$ . Although  $\langle \text{LAT}_i \rangle$  are discrete vocabulary tokens, their contextualized final-layer hidden states  $\mathbf{h}_{lat}^i$  serve as continuous latent reasoning representations. During autoregressive generation, each latent state attends to the image, the query, and all preceding latent-token positions. Consequently, the final state  $\mathbf{h}_{lat}^K$  aggregates information from the entire latent-token sequence. The subsequently generated description  $\mathcal{D}$  and answer  $\mathcal{A}$  can attend to all preceding latent states through causal self-attention. At segmentation time, we provide SAM3 with the text prompt derived from  $\mathcal{D}$  and the bounding-box prompt extracted from  $\mathcal{A}$  to obtain the final mask.

## Two-Stage Training Pipeline

We train LIRSeg in two stages. Stage 1 spatially grounds the latent tokens, providing a stable initialization for reinforcement learning. Stage 2 then optimizes the policy MLLM and the latent tokens with GRPO under segmentation rewards.

**Stage 1: Warm Start via Spatial Alignment.** Directly applying RL to randomly initialized latent tokens is unstable, as sparse segmentation rewards provide insufficient supervision for learning meaningful latent tokens from scratch. To this end, we introduce latent spatial alignment, which explicitly grounds latent tokens in target-relevant visual regions without distilling verbose teacher-generated reasoning traces.

Let  $\mathbf{h}_{lat}^K \in \mathbb{R}^d$  denote the hidden state of the  $K$ -th latent token at the last transformer layer of the MLLM, and let

$P \in \mathbb{R}^{N \times d}$  denote the visual patch features produced by the vision encoder. We compute a spatial similarity map as

$$s = \mathbf{h}_{lat}^K P^\top, \quad S_{grid} = \text{reshape}(s) \in \mathbb{R}^{H \times W}, \quad (1)$$

where  $S_{grid}$  indicates the visual regions associated with the latent representation. We supervise  $S_{grid}$  with the ground-truth mask in two directions. Token-to-mask (T2M) alignment upsamples  $S_{grid}$  to the pixel resolution and matches it to the mask, promoting boundary localization. Conversely, mask-to-token (M2T) alignment downsamples the Gaussian-smoothed mask into a token-level target and aligns it with  $S_{grid}$ , injecting spatial semantics into the latent representation. Both directions are optimized with a combination of binary cross-entropy and Dice losses:

$$\mathcal{L}_{T2M} = \text{BCE}(\text{up}(S_{grid}), M) + \text{Dice}(\cdot), \quad (2)$$

$$\mathcal{L}_{M2T} = \text{BCE}(S_{grid}, \text{down}(G_\sigma * M)) + \text{Dice}(\cdot). \quad (3)$$

The total loss of Stage 1 is

$$\mathcal{L}_{spatial} = \mathcal{L}_{T2M} + \mathcal{L}_{M2T}. \quad (4)$$

Only the last transformer layer of the MLLM and the learnable latent-token embeddings are updated; all remaining backbone parameters are frozen. This stage provides the latent representation with a spatially meaningful initialization, allowing Stage 2 to refine object-grounded tokens rather than optimize from random noise.

**Stage 2: RL and Reward Design.** In Stage 2, we optimize the policy using GRPO (Shao et al. 2024). Given an input  $x = (I, Q)$ , we sample a group of  $n$  responses  $\{y_1, \dots, y_n\} \sim \pi_{\theta_{old}}$  and assign each response a reward  $r_i$ .

The group-relative advantage is computed as

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}, \quad \bar{r} = \frac{1}{n} \sum_j r_j. \quad (5)$$

where  $\sigma_r$  is the standard deviation within the group.

The reward function jointly optimizes segmentation quality, localization accuracy, conciseness, and format:

$$r = r_{\text{mask}} + r_{\text{bbox}} + r_{\text{len}} + r_{\text{fmt}}. \quad (6)$$

The first two terms provide task-specific supervision. To encourage latent reasoning to encode localization cues without relying on bounding-box coordinates, we evaluate the target description  $\mathcal{D}$  by its ability to guide a segmentation model:

$$r_{\text{mask}} = \text{IoU}(\text{Mask}(I, \mathcal{D}), M_{\text{gt}}), \quad (7)$$

where  $\text{Mask}(I, \mathcal{D})$  is the predicted mask and  $M_{\text{gt}}$  is the ground-truth mask. Direct localization is supervised using standard bounding-box IoU:

$$r_{\text{bbox}} = \text{IoU}(B_{\text{pred}}, B_{\text{gt}}). \quad (8)$$

The remaining terms prevent overly verbose or malformed outputs. We reward concise descriptions with

$$r_{\text{len}} = \mathbf{1}_{[1,4]}(|\mathcal{D}|), \quad (9)$$

where  $|\mathcal{D}|$  denotes the word count. We also introduce a format reward  $r_{\text{fmt}}$  to check whether the output conforms to the predefined format  $\mathcal{F}_{\text{lat}}$ .

### Optimization for Learnable Latent Tokens

**Signal: Extreme Advantage Sampling** Replacing verbose CoT with latent tokens reduces rollout diversity because these tokens are insufficiently pretrained to support diverse reasoning trajectories. Consequently, rewards become tightly clustered and advantages concentrate near zero, yielding weak and noise-sensitive policy-gradient signals. We therefore discard the near-zero middle and estimate the gradient using only the two tails:

$$\hat{g}_{\text{ext}} = \frac{1}{K_h + K_l} \sum_{i=1}^n m_i A_i \nabla_{\theta} \log \pi_{\theta}(y_i | x), \quad (10)$$

$m_i = \mathbf{1}[A_i \leq A_{(K_l)} \vee A_i \geq A_{(n-K_h+1)}]$ , where  $A_{(j)}$  is the  $j$ -th order statistic. Thus, near-zero advantages yield no update, while extreme advantages give clear directions to reinforce or suppress behaviors.

### Utilization: Decoupled Exploration–Stability Update

The generated output tokens serve two complementary roles in  $\mathcal{D}$  and  $\mathcal{A}$ . Tokens in  $\mathcal{D}$  and  $\mathcal{A}$  guide the model’s exploration toward the target and are optimized by the IoU reward, inducing an exploration gradient  $\mathbf{g}_{\text{exp}}$ . Structural tokens, such as the `<answer>` tag, are optimized by a format reward and induce a stability gradient  $\mathbf{g}_{\text{sta}}$ . These gradients exhibit negative cosine similarity, such that naive aggregation can attenuate the reward-aligned exploration signal.

We address this conflict with a decoupled update: we preserve  $\mathbf{g}_{\text{exp}}$  and project  $\mathbf{g}_{\text{sta}}$  onto the subspace orthogonal to the exploration direction:

$$\mathbf{g}_{\text{sta}}^{\perp} = \mathbf{g}_{\text{sta}} - \frac{\langle \mathbf{g}_{\text{sta}}, \mathbf{g}_{\text{exp}} \rangle}{\|\mathbf{g}_{\text{exp}}\|^2 + \epsilon} \mathbf{g}_{\text{exp}}. \quad (11)$$

The final update is  $\mathbf{g} = \mathbf{g}_{\text{exp}} + \mathbf{g}_{\text{sta}}^{\perp}$ . By construction, the cosine similarity  $\langle \mathbf{g}_{\text{exp}}, \mathbf{g}_{\text{sta}}^{\perp} \rangle = 0$ . Thus, the reward-aligned exploration gradient is well-preserved, while the stability gradient contributes only through orthogonal, non-interfering directions that maintain output structure.

**Capacity: Latent Diversity Amplification** Although the  $K$  registered special tokens have distinct token identities, their contextual hidden states may collapse toward similar directions in the absence of explicit regularization. Such collapse makes the latent representation redundant and limits the information available for predicting  $Y = (\mathcal{D}, \mathcal{A})$ . Let  $z_i := \mathbf{h}_{\text{lat}}^i$  denote the final-layer contextual hidden state associated with  $\langle \text{LAT}_i \rangle$ , and define  $\mathcal{Z} = \{z_1, \dots, z_K\}$ . Their mean pairwise cosine similarity as

$$\bar{s} = \frac{2}{K(K-1)} \sum \cos(z_i, z_j), \quad (12)$$

and introduce the diversity regularizer:

$$\mathcal{L}_{\text{div}} = \text{softplus}(\bar{s} - \tau_{\text{div}}), \quad (13)$$

where  $\tau_{\text{div}}$  is a similarity threshold. The penalty remains weak when the latent states are sufficiently dispersed and grows approximately linearly as they become aligned. By penalizing redundant latent directions,  $\mathcal{L}_{\text{div}}$  helps preserve representational rank and encourages different tokens to capture complementary aspects of the reasoning process.

### Theoretical Motivation

**Verbalization Bottleneck.** LIRSeg retains reasoning in continuous latent states  $\mathcal{Z}$  instead of verbalizing it into a discrete token sequence  $\mathcal{T}$  before mask construction. Consider matched upstream computation in which  $\mathcal{T} \sim q(\mathcal{T} | \mathcal{Z})$  is the reasoning interface passed downstream. Then  $M - \mathcal{Z} - \mathcal{T}$  forms a Markov chain: given the latent state, the text tokens are conditionally independent of the target mask. The data-processing inequality gives  $I(M; \mathcal{T}) \leq I(M; \mathcal{Z})$ , or equivalently

$$I(M; \mathcal{Z}) - I(M; \mathcal{T}) = I(M; \mathcal{Z} | \mathcal{T}) \geq 0. \quad (14)$$

Verbalization is therefore lossless only when  $\mathcal{T}$  is a sufficient statistic of  $\mathcal{Z}$  for predicting  $M$ . Otherwise, task-relevant information in the latent state is discarded by the textual interface. This result does not assume that continuous representations are intrinsically more informative; it only motivates preserving the latent reasoning state rather than forcing an intermediate verbalization.

**Latent sufficiency.** Because LIRSeg decodes the mask directly from  $\mathcal{Z}$ , the residual uncertainty is  $H(M | \mathcal{Z})$ . Had one instead used the textual interface, the chain rule for entropy together with the Markov property  $H(M | \mathcal{Z}, \mathcal{T}) = H(M | \mathcal{Z})$  yields

$$H(M | \mathcal{T}) = H(M | \mathcal{Z}) + I(M; \mathcal{Z} | \mathcal{T}), \quad (15)$$

which implies  $H(M | \mathcal{T}) \geq H(M | \mathcal{Z})$ . The inequality is strict whenever  $\mathcal{T}$  is not sufficient for  $\mathcal{Z}$ . Thus the latent state is never less informative than its verbalization; preserving  $\mathcal{Z}$  avoids the additional prediction uncertainty introduced by discrete compression.

Table 1: **Performance comparison on ReasonSeg.** We report reasoning tokens to measure efficiency. \* marks models trained on the VisionReasoner dataset. Bold and underlined denote best and second-best results, respectively.

| Method                | Venue      | Language Model | Seg. | ReasonSeg Val |             |             | ReasonSeg Test |             |             |
|-----------------------|------------|----------------|------|---------------|-------------|-------------|----------------|-------------|-------------|
|                       |            |                |      | Tokens ↓      | gIoU ↑      | cIoU ↑      | Tokens ↓       | gIoU ↑      | cIoU ↑      |
| OVSeg                 | CVPR’23    | CLIP ViT-L     | –    | –             | 28.5        | 18.6        | –              | 26.1        | 20.8        |
| ReLA                  | CVPR’23    | BERT           | –    | –             | 22.4        | 19.9        | –              | 21.3        | 22.0        |
| LISA                  | CVPR’24    | LLaVA1.5-7B    | SAM1 | –             | 61.3        | 62.9        | –              | 55.6        | 56.9        |
| CoReS                 | ECCV’24    | LLaVA-7B       | SAM1 | –             | 59.4        | –           | –              | 52.4        | –           |
| Seg-Zero              | arxiv      | Qwen2.5-VL-7B  | SAM2 | 90.7          | 61.6        | 52.5        | 90.6           | 58.2        | 52.3        |
| SAM-R1                | NeurIPS’25 | Qwen2.5-VL-7B  | SAM2 | –             | 64.0        | 55.8        | –              | 60.2        | 54.3        |
| PixelThink            | ICLR’26    | Qwen2.5-VL-7B  | SAM2 | 46.9          | 63.8        | 62.6        | 47.6           | 60.1        | 55.7        |
| VisionReasoner*       | ICLR’26    | Qwen2.5-VL-7B  | SAM3 | 80.8          | 66.3        | 59.8        | 84.8           | 63.6        | 58.2        |
| VisionReasoner        | ICLR’26    | Qwen2.5-VL-7B  | SAM2 | 85.3          | 65.4        | 60.3        | 81.4           | 62.3        | 54.6        |
| VisionReasoner        | ICLR’26    | Qwen2.5-VL-7B  | SAM3 | 64.8          | 65.8        | 61.5        | 64.7           | 65.5        | 59.2        |
| DR <sup>2</sup> Seg*  | ICML’26    | Qwen2.5-VL-7B  | SAM2 | 46.2          | 67.5        | 60.0        | 55.4           | 64.8        | 62.8        |
| DR <sup>2</sup> Seg   | ICML’26    | Qwen2.5-VL-7B  | SAM2 | <u>26.9</u>   | 68.5        | 65.8        | <u>27.2</u>    | 66.1        | <u>63.6</u> |
| DR <sup>2</sup> Seg   | ICML’26    | Qwen2.5-VL-7B  | SAM3 | 31.2          | <u>69.4</u> | <u>66.4</u> | 30.7           | <u>66.5</u> | 61.7        |
| <b>LIRSeg (Ours)*</b> | –          | Qwen2.5-VL-7B  | SAM3 | <b>4</b>      | 69.0        | 60.6        | <b>4</b>       | 65.2        | 63.1        |
| <b>LIRSeg (Ours)</b>  | –          | Qwen2.5-VL-7B  | SAM3 | <b>4</b>      | <b>71.7</b> | <b>67.7</b> | <b>4</b>       | <b>70.4</b> | <b>63.8</b> |

**Compact attention.** Explicit CoT expands to a long sequence  $\mathcal{T}$  with  $|\mathcal{T}| \gg |\mathcal{Z}|$ , whose tokens compete with visual features for attention capacity. Combining Eq. (14) with the token budget gives the per-token task information

$$\frac{I(M; \mathcal{Z})}{|\mathcal{Z}|} \gg \frac{I(M; \mathcal{T})}{|\mathcal{T}|}. \quad (16)$$

Because  $\mathcal{Z}$  is compact and interacts directly with visual features without intermediate language, it avoids the attention dilution and increased effective distance between perception tokens caused by redundant text. Each latent token therefore participates in undiluted visual attention, whereas language tokens insert non-visual keys and values that lengthen the reasoning path and introduce irrelevant image distractors.

## Experiments

### Experimental Settings

**Datasets.** For training, we use either the ReasonSeg training set (only 239 samples) or the VisionReasoner-7K dataset separately, to examine generalization under different training data. We evaluate on three reasoning segmentation benchmarks and the RefCOCO referring suite. ReasonSeg tests multi-instance reasoning, MUSE tests multi-target semantic discrimination, MMR requires joint object- and part-level masks, and RefCOCO verifies generalization.

**Evaluation Metrics.** Following prior work (Lai et al. 2024; He et al. 2026), we report gIoU and cIoU for accuracy, and the average number of reasoning tokens for efficiency.

**Experimental Details.** All experiments are conducted on 4 H100 GPUs with a per-device batch size of 8. By default we use Qwen2.5-VL-7B (Bai et al. 2025) and SAM3 (Carion et al. 2025). We set  $K = 4$  latent tokens. Stage 1 performs SFT for 500 steps. Stage 2 applies RL for 5 epochs with  $n = 16$  rollouts,  $K_h = K_l = 4$  extreme sampling.

## Main Results

**Comparison Methods.** We compare LIRSeg with non-MLLM segmentation methods (Liang et al. 2023; Liu, Ding, and Jiang 2023b), SFT-based MLLM methods (Lai et al. 2024; Bao et al. 2024; Xia et al. 2024; Rasheed et al. 2024; Jang et al. 2025), explicit-CoT RL methods (Liu et al. 2025a; Zhu et al. 2025a; Huang et al. 2025; Wang et al. 2025; Liu et al. 2025b; He et al. 2026). Finally, our LIRSeg, a latent reasoning method, replaces explicit CoT with a compact set of learnable latent tokens, enabling efficient encoding of task-relevant visual and semantic information.

### Reasoning Segmentation Results

**Multi-Target Single-Semantic Scenes (ReasonSeg).** Table 1 reports results on ReasonSeg. Under the same Qwen2.5-VL-7B backbone and SAM3 segmenter, LIRSeg achieves 71.7/67.7 gIoU/cIoU on val and 70.4/63.8 on test, surpassing VisionReasoner by 5.9/6.2 and 4.9/4.6, respectively. LIRSeg reduces reasoning tokens by 16× (64.8 → 4). Per-sample inference time drops from 0.783 s to 0.512 s, achieving 53% speedup over VisionReasoner. It also outperforms the recent state-of-the-art method DR<sup>2</sup>Seg by 2.3 and 3.9 gIoU on val/test. These results demonstrate that latent reasoning effectively alleviates attention interference, improving both accuracy and efficiency. Latent reasoning also generalizes well in the zero-shot setting: LIRSeg\*, trained only on the VisionReasoner-7K dataset (short referring expressions without complex reasoning) and directly evaluated on ReasonSeg, improves over VisionReasoner\* by 2.7 gIoU on val while reducing tokens by 20× (80.8 → 4). This suggests that latent reasoning is not task-specific overfitting; rather, its compact bottleneck directly encodes task-relevant visual evidence, enabling transfer from simple to complex queries.

**Multi-Target Multi-Semantic Scenes (MUSE).** Table 2 (left) reports MUSE results, where each query describes



Figure 3: **Qualitative comparisons between VisionReasoner and our LIRSeg.** The representative samples are selected from simple single-object to complex multi-object scenarios, which reveals severe attention interference.

Table 2: **Results on MUSE and MMR benchmarks.** Performance is reported in terms of gIoU. A Dash indicates that the corresponding method did not report results.

| Method            | MUSE        |             | MMR         |             |
|-------------------|-------------|-------------|-------------|-------------|
|                   | val         | test        | val         | test        |
| LISA              | 42.0        | 38.9        | 19.4        | 19.5        |
| PixelLM           | 42.6        | 39.2        | –           | –           |
| POPEN             | 45.4        | 42.4        | –           | –           |
| GSVA              | –           | –           | 19.8        | 21.2        |
| GLaMM             | –           | –           | 26.9        | 30.3        |
| M <sup>2</sup> SA | –           | –           | <u>27.8</u> | <u>30.9</u> |
| VisionReasoner    | 50.5        | 47.7        | 26.7        | 28.4        |
| LIRSeg            | <b>55.8</b> | <b>54.8</b> | <b>29.6</b> | <b>33.1</b> |

multiple targets with diverse semantics. LIRSeg outperforms VisionReasoner by 5.3 and 7.1 gIoU on val and test splits, respectively. These gains show that latent tokens effectively encode multi-target semantics while avoiding the bottleneck and attention interference of explicit text reasoning.

**Multi-Target Multi-Granularity Scenes (MMR).** Table 2 reports results on MMR, which requires segmentation at both the object and part levels. LIRSeg outperforms VisionReasoner by 2.9 gIoU on the val split and 4.7 gIoU on the test split. The consistent gains demonstrate that latent reasoning remains effective across varying levels of granularity.

**Referring Segmentation Results** Table 3 evaluates generalization on the RefCOCO series, on which LIRSeg achieves

Table 3: **Referring segmentation, RefCOCO series.**

| Method              | RefCOCO<br>testA | RefCOCO+<br>testA | RefCOCOg<br>test |
|---------------------|------------------|-------------------|------------------|
| LISA                | 76.5             | 67.4              | 68.5             |
| Seg-Zero            | <b>80.3</b>      | <b>76.2</b>       | <u>72.6</u>      |
| VisionReasoner      | <u>78.8</u>      | 75.1              | 71.5             |
| DR <sup>2</sup> Seg | 78.7             | <u>75.4</u>       | 72.2             |
| LIRSeg              | 78.7             | 75.0              | <b>72.7</b>      |

78.7/75.0/72.7 gIoU, attaining the best result on RefCOCOg while remaining competitive on the others. This demonstrates that latent reasoning also preserves strong performance in simple referring scenarios.

**Qualitative Results** Figure 3 provides a qualitative comparison between VisionReasoner and LIRSeg. LIRSeg produces concise, target-specific outputs and consistently localizes the intended object, even in cluttered scenes. In contrast, the explicit-reasoning baseline is more prone to drifting toward distractors mentioned in its CoT traces.

## Diagnostic Experiments

**Ablation on Core Components** Table 4 shows leave-one-out ablations on ReasonSeg validation set, where removing any component consistently drops gIoU by 2–3 points. In particular, ablating LDA causes the largest decline (68.2 gIoU), and the consistent degradation across all variants clearly underscores their complementary synergy in boosting both seg-

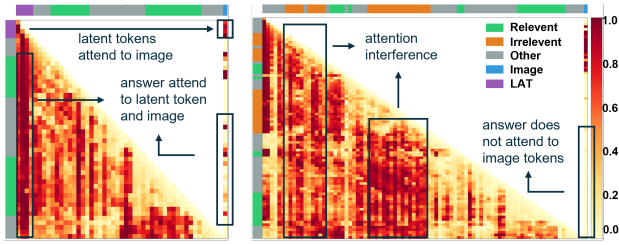


Figure 4: **Attention confusion phenomenon.** Left: latent reasoning. Right: explicit reasoning.

Table 4: **Component ablation of LIRSeg.** EAS: extreme advantage sampling; DES: decoupled exploration–stability update; LDA: latent diversity amplification.

| Configuration         | Tokens ↓ | gIoU ↑      | cIoU ↑      |
|-----------------------|----------|-------------|-------------|
| <b>LIRSeg (full)</b>  | 4        | <b>71.7</b> | <b>67.7</b> |
| w/o Stage 1 alignment | 4        | 68.6        | 60.9        |
| w/o EAS               | 4        | 69.2        | 61.0        |
| w/o DES               | 4        | 69.9        | 59.4        |
| w/o LDA               | 4        | 68.2        | 60.1        |

mentation accuracy and robustness, further confirming that each component targets a distinct informational bottleneck.

**Empirical Analysis of Attention Interference** Figure 4 compares attention patterns under explicit and latent reasoning. In the explicit-reasoning model (right), the output token attends broadly over the generated reasoning chain, assigning attention to irrelevant text tokens (orange) while weakening its focus on image tokens (blue) and relevant text (green). This pattern illustrates attention interference, where redundant verbal CoT disrupts visual grounding. In contrast, the latent-reasoning model (left) concentrates latent-token attention (purple) on task-relevant image regions. The answer token then attends directly to the image and latent tokens, avoiding interference from irrelevant textual reasoning.

Table 5: **Stratified comparison by reasoning difficulty.** Samples are split at the median think-token count of the explicit baseline.  $\Delta$ gIoU denotes the absolute improvement of LIRSeg over the explicit baseline.

| Group   | Threshold | Samples | Explicit | Latent | $\Delta$ gIoU |
|---------|-----------|---------|----------|--------|---------------|
| Overall | –         | 200     | 65.8     | 71.7   | +5.9          |
| Easy    | $\leq 68$ | 102     | 67.5     | 70.6   | +3.1          |
| Hard    | $> 68$    | 98      | 64.0     | 72.9   | <b>+8.9</b>   |

**Impact on Easy and Hard Reasoning Cases** We use the median reasoning tokens from VisionReasoner to divide samples into easy and hard groups. As shown in Table 5, LIRSeg improves gIoU by 8.9 on hard samples, compared to 3.1 on easy samples. This larger improvement suggests that latent reasoning goes beyond compact visual anchoring by capturing task-relevant compositional information whose importance increases with reasoning complexity. The results

also support the attention-interference analysis in Figure 4, indicating that removing explicit reasoning is particularly beneficial for queries requiring longer inference chains.

Table 6: **Ablation on Number of latent tokens.**

| $K$ | gIoU ↑      | cIoU ↑      |
|-----|-------------|-------------|
| 0   | 67.6        | 55.4        |
| 1   | 69.1        | 59.8        |
| 2   | 69.5        | 64.9        |
| 4   | <b>71.7</b> | <b>67.7</b> |
| 8   | 68.1        | 58.9        |

**Effect of Latent Token Number** Table 6 ablates the latent token count  $K$ . A single token ( $K = 1$ ) underperforms 69.1 gIoU due to insufficient capacity, while  $K = 8$  introduces redundancy and drops to 68.1 gIoU. The best result is at  $K = 4$ . This confirms a capacity–compression trade-off: too few tokens fail to encode sufficient reasoning, while too many weaken the bottleneck and encourage redundancy.

Table 7: **Generality across backbones and segmenters.**

| MLLM       | Seg. | Val         |             | Test        |             |
|------------|------|-------------|-------------|-------------|-------------|
|            |      | gIoU        | cIoU        | gIoU        | cIoU        |
| Qwen2.5-VL | SAM2 | 68.8        | 62.6        | 67.8        | 64.5        |
| Qwen2.5-VL | SAM3 | 71.7        | 67.7        | 70.4        | 63.8        |
| Qwen3-VL   | SAM2 | <b>73.2</b> | <b>67.9</b> | <b>71.0</b> | <b>68.7</b> |
| Qwen3-VL   | SAM3 | 72.1        | 62.8        | 70.8        | 64.4        |

**Generality across Backbones and Segmenters** Table 7 evaluates LIRSeg across different MLLM backbones and segmenters. All configurations outperform their explicit-reasoning counterparts on ReasonSeg, with Qwen3-VL (Yang and Team 2025) + SAM2 achieving the best result of 73.2/71.0 gIoU on the val/test splits. These results demonstrate the strong architectural generalization of LIRSeg.

## Conclusion

LIRSeg introduces latent reasoning for reasoning segmentation by replacing explicit CoT with a compact set of learnable latent tokens. These tokens are first grounded in task-relevant visual evidence through spatial alignment and are then optimized with segmentation rewards. To make this compact latent bottleneck effective, we develop three complementary mechanisms: extreme-advantage sampling to strengthen reward signals, decoupled exploration–stability updates to preserve useful gradients, and latent diversity amplification to prevent representational collapse. Across multiple MLLM backbones and segmenters, LIRSeg consistently improves segmentation accuracy while reducing reasoning tokens by approximately  $16\times$ , achieving state-of-the-art results on ReasonSeg, MUSE, and MMR and remaining competitive on RefCOCO. These results show that explicit verbal reasoning is not necessary for accurate reasoning segmentation and provide new insights into how compact latent computation can improve both visual grounding and reasoning efficiency.

## References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Bao, X.; Sun, S.; Ma, S.; Zheng, K.; Guo, Y.; Zhao, G.; Zheng, Y.; and Wang, X. 2024. CoReS: Orchestrating the Dance of Reasoning and Segmentation. In *European Conference on Computer Vision*, 187–204. Springer.
- Carion, N.; Gustafson, L.; Hu, Y.-T.; Debnath, S.; Hu, R.; Suris, D.; Ryali, C.; Alwala, K. V.; Khedr, H.; Huang, A.; et al. 2025. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719*.
- Deng, Y.; Choi, Y.; and Shieber, S. M. 2024. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step. *arXiv preprint arXiv:2405.14838*.
- Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16321–16330.
- Goyal, S.; Ji, Z.; Rawat, A. S.; Menon, A. K.; Kumar, S.; and Nagarajan, V. 2024. Think Before You Speak: Training Language Models with Pause Tokens. In *The Twelfth International Conference on Learning Representations*.
- Hao, S.; Sukhbaatar, S.; Su, D.; Li, X.; Hu, Z.; Weston, J.; and Tian, Y. 2024. Training Large Language Models to Reason in a Continuous Latent Space. *arXiv preprint arXiv:2412.06769*.
- He, Y.; Chen, W.; Jian, Z.; Guo, T.; Zhou, W.; Li, M.; Yang, S.; and Yang, W. 2026. DR<sup>2</sup>Seg: Decomposed Two-Stage Rollouts for Efficient Reasoning Segmentation in Multimodal Large Language Models. In *International Conference on Machine Learning (ICML 2026)*. Seoul, South Korea: PMLR.
- Huang, J.; Xu, Z.; Zhou, J.; Liu, T.; Xiao, Y.; Ou, M.; Ji, B.; Li, X.; and Yuan, K. 2025. SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning. *Advances in Neural Information Processing Systems*.
- Jang, D.; Cho, Y.; Lee, S.; Shin, T.; and Kim, D.-S. 2025. MMR: A Large-Scale Benchmark Dataset for Multi-Target and Multi-Granularity Reasoning Segmentation. *arXiv preprint arXiv:2503.13881*.
- Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.
- Li, B.; Sun, X.; Liu, J.; Wang, Z.; Wu, J.; Yu, X.; Chen, H.; Barsoum, E.; Chen, M.; and Liu, Z. 2025. Latent Visual Reasoning. *arXiv preprint arXiv:2509.24251*.
- Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-Vocabulary Semantic Segmentation with Mask-Adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.
- Liu, C.; Ding, H.; and Jiang, X. 2023a. GRES: Generalized Referring Expression Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23592–23601.
- Liu, C.; Ding, H.; and Jiang, X. 2023b. GRES: Generalized Referring Expression Segmentation. *arXiv:2306.00968*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.
- 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.; Qu, T.; Zhong, Z.; Peng, B.; Liu, S.; Yu, B.; and Jia, J. 2025b. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.12081*.
- Lu, Y.; Cao, J.; Wu, Y.; Li, B.; Tang, L.; Ji, Y.; Wu, C.; Wu, J.; and Zhu, W. 2025. RSVP: Reasoning Segmentation via Visual Prompting and Multi-modal Chain-of-Thought. *arXiv:2506.04277*.
- Pham, T.-H.; and Ngo, C. 2025. Multimodal Chain of Continuous Thought for Latent-Space Reasoning in Vision-Language Models. *arXiv preprint arXiv:2508.12587*.
- Qian, R.; Deng, C.; Huang, Q.; Xiong, J.; Li, M.; Zhou, Y.; Zhai, W.; Chen, J.; and Dou, D. 2026. AnchorSeg: Language Grounded Query Banks for Reasoning Segmentation. *arXiv:2604.18562*.
- Rasheed, H.; Maaz, M.; Mullappilly, S. S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. *arXiv:2311.03356*.
- Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. PixelLM: Pixel Reasoning with Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.
- 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*.
- Wang, J.; and Ke, L. 2024. LLM-Seg: Bridging Image Segmentation and Large Language Model Reasoning. *arXiv:2404.08767*.
- Wang, S.; Fang, G.; Kong, L.; Li, X.; Xu, J.; Yang, S.; Li, Q.; Zhu, J.; and Wang, X. 2025. PixelThink: Towards Efficient Chain-of-Pixel Reasoning. *arXiv preprint arXiv:2505.23727*.
- Wang, X.; Zhang, S.; Li, S.; Kallidromitis, K.; Li, K.; Kato, Y.; Kozuka, K.; and Darrell, T. 2024. SegLLM: Multi-round Reasoning Segmentation. *arXiv:2410.18923*.
- Wei, C.; Zhong, Y.; Tan, H.; Zeng, Y.; Liu, Y.; Zhao, Z.; and Yang, Y. 2024. InstructSeg: Unifying Instructed Visual Segmentation with Multi-modal Large Language Models. *arXiv:2412.14006*.
- Wu, T.-H.; Biamby, G.; Chan, D.; Dunlap, L.; Gupta, R.; Wang, X.; Gonzalez, J. E.; and Darrell, T. 2023. See, Say, and Segment: Teaching LMMs to Overcome False Premises. *arXiv:2312.08366*.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. *arXiv:2312.10103*.

Yang, A.; and Team, Q. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2024. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. *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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18155–18165.

Yang, Z.; Yu, X.; Chen, D.; Shen, M.; and Gan, C. 2025. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens. *arXiv preprint arXiv:2506.17218*.

You, Z.; and Wu, Z. 2025. Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning. *arXiv:2506.22624*.

Zelikman, E.; Harik, G.; Shao, Y.; Jayasiri, V.; Haber, N.; and Goodman, N. D. 2024. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. *arXiv preprint arXiv:2403.09629*.

Zhang, Y.; Ma, Z.; Gao, X.; Shakiah, S.; Gao, Q.; and Chai, J. 2024. GROUNDHOG: Grounding Large Language Models to Holistic Segmentation. *arXiv:2402.16846*.

Zhao, Y.; Pang, Y.; Zuo, J.; Ji, W.; Zhou, K.; Fan, B.; Cao, Y.; Zhang, L.; Liu, X.; Lu, H.; Lin, W.; Tao, D.; and Zhao, X. 2026. ConceptSeg-R1: Segment Any Concept via Meta-Reinforcement Learning. *arXiv:2605.20385*.

Zhu, L.; Chen, T.; Xu, Q.; Liu, X.; Ji, D.; Wu, H.; Soh, D. W.; and Liu, J. 2025a. POPEN: Preference-Based Optimization and Ensemble for LVLM-Based Reasoning Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30231–30240.

Zhu, L.; Ouyang, B.; Zhang, Y.; Cheng, T.; Hu, R.; Shen, H.; Ran, L.; Chen, X.; Yu, L.; Liu, W.; and Wang, X. 2025b. LENS: Learning to Segment Anything with Unified Reinforced Reasoning. *arXiv:2508.14153*.