PARS: Unifying Reasoning and Perceptual Reinforcement for Multimodal Segmentation

Hikaru Tanabe[1], Yuta Baba[1], and Keiji Yanai[1]

The University of Electro-Communications, Tokyo, Japan
{tanabe-h,baba-y}@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. Reasoning segmentation aims to localize target regions at pixel-level granularity given an image and a complex text query that requires advanced reasoning to identify the target object. Existing approaches that apply reinforcement learning to Multimodal Large Language Models (MLLMs) largely rely on naive adaptations of GRPO, resulting in suboptimal learning efficiency and limited improvement in perceptual capability, which is critical for accurate visual grounding in reasoning segmentation. To address these limitations, we propose PARS, a unified framework that explicitly integrates perceptual reinforcement into reasoning segmentation. Within this framework, we introduce AP2O, a novel group-based policy optimization algorithm that enables stable co-optimization of reasoning and perception. Specifically, AP2O employs a soft-gated importance ratio to stabilize policy updates for reasoning optimization, while introducing a Low-Frequency Patch Perception Loss to effectively enhance perceptual consistency. Extensive experiments demonstrate that PARS significantly outperforms the existing RL baseline Seg-Zero at comparable model scales, achieving gains of +5.5 gIoU on the ReasonSeg validation set and +4.5 cIoU on the test set.

Keywords: Reasoning Segmentation · Multimodal Large Language Models · Reinforcement Learning

1 Introduction

Reasoning segmentation [10] is a task that aims to estimate the target region at pixel-level granularity given an image and a text description in which the target object is implicitly specified (e.g., "an ingredient that adds sourness to a dish"). In contrast, conventional Referring Expression Segmentation (RES) primarily considers text that explicitly names the target object (e.g., "lemon"). Reasoning segmentation is more challenging in that it requires the model to correctly interpret the textual description and identify the underlying target object. Moreover, because the task requires pixel-level localization, the model must also precisely perceive and distinguish visual attributes of objects in the image in order to ground the identified target. Consequently, a reasoning segmentation model must possess both strong perceptual capability for visual grounding and robust reasoning ability to interpret the instruction.

Looking ahead to scenarios in which diverse users issue instructions to autonomous embodied systems, users are more likely to provide concise, goal-oriented commands (e.g., "change the TV channel") rather than explicitly decomposing the required steps (e.g., "go near the table, find the remote control, and press the button to change the channel"). Therefore, the capability of a system to properly understand user instructions and the surrounding environment, and to autonomously infer and ground the necessary objects for goal achievement, is crucial for expanding the applicability of embodied systems. Such capability is particularly important in a wide range of domains, including robot control and automatic graphic design editing, where versatility and usability are essential. In this context, reasoning segmentation serves as an important testbed for evaluating both perceptual accuracy in understanding the visual world and reasoning ability in interpreting user instructions. For these reasons, the task holds substantial social significance for future AI applications.

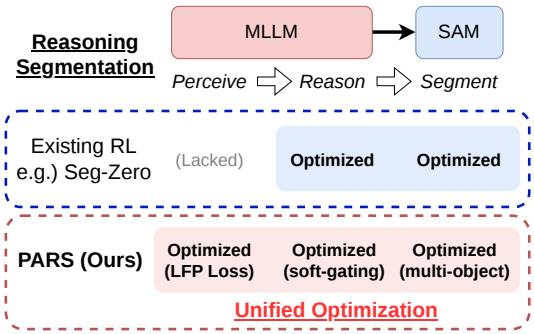


Fig. 1: Perceptual failures propagate through reasoning and cause large segmentation errors in reasoning segmentation. PARS addresses this problem through a unified reinforcement learning framework that jointly optimizes perception and reasoning. The key idea is to stabilize reasoning optimization via soft-gated policy updates while explicitly enhancing visual perception through perceptual losses, enabling perception-aware reasoning segmentation.

Despite its importance, most existing approaches to reasoning segmentation rely on Supervised Fine-Tuning (SFT). However, SFT-based models tend to suffer from degraded performance under out-of-distribution (OOD) settings.

Recently, Reinforcement Learning with Verifiable Rewards (RLVR) [11] has been reported to effectively enhance the reasoning capabilities of Large Language Models (LLMs) [7, 28]. Motivated by these findings, recent reasoning segmentation studies have introduced RLVR frameworks targeting Multimodal Large Language Models (MLLMs), which are used as reasoning modules, to improve reasoning ability and enhance OOD generalization.

Seg-Zero [18] introduced an RLVR framework based on group-based policy optimization using GRPO [28] into reasoning segmentation, improving OOD segmentation accuracy by enhancing reasoning performance. However, the adopted training strategy inherits stability issues from GRPO, and improvements at the objective level have not been thoroughly investigated. Moreover, improving perceptual capability regarding object type and location is a promising direction for advancing reasoning segmentation. Nevertheless, existing group-based policy optimization approaches for this task have not explored systematic perceptual enhancement.

The objective of this study is to improve group-based policy optimization for reasoning segmentation in order to enhance region estimation accuracy. To this end, we propose PARS: Perception-Aware Reasoning Segmentation, a unified framework that integrates perceptual enhancement into RL-based reasoning segmentation (Fig. 1). Within this framework, we introduce Adaptive Perception-Aware Policy Optimization (AP2O), a novel group-based policy optimization algorithm designed to stably co-optimize reasoning and perceptual capabilities for accurate visual grounding.

First, to improve reasoning performance and training stability in conventional policy optimization, we introduce several modifications at the objective level, including a soft-gated importance ratio, the removal of the KL regularization term, and dynamic sampling. These modifications effectively enhance the model's capability to identify target objects through reasoning, thereby improving segmentation accuracy.

Second, focusing on perceptual enhancement that has been underexplored in prior work, we introduce a perception loss into reasoning segmentation, motivated by reports that perceptual losses can stably improve perceptual capability in multimodal RLVR. This encourages stable improvements in recognizing object type and location, leading to better visual grounding performance.

Furthermore, to address the issue that conventional perception losses may reduce changes in the policy output distribution, we propose a novel Low-Frequency Patch Perception Loss. This loss promotes effective masking of semantic information during perceptual editing of visual inputs, thereby strengthening the perceptual improvement effect.

By integrating these techniques, we achieve substantial performance gains on ReasonSeg, a standard benchmark for reasoning segmentation. Compared with the prior method Seg-Zero trained with the same base model, our approach improves validation gIoU by +5.5 and test cIoU by +4.5, outperforming existing methods at comparable model scales.

Our main contributions are summarized as follows:

- We propose PARS, a unified framework that incorporates explicit perceptual reinforcement into RL-based reasoning segmentation, enabling effective joint optimization of reasoning and pixel-level perception.
- We introduce AP2O, a novel group-based policy optimization algorithm that improves both training stability and efficiency, employing a soft-gated importance ratio for stable reasoning updates and a Low-Frequency Patch Perception Loss to enhance perceptual consistency.
- We achieve substantial and consistent performance gains over the RL baseline Seg-Zero at comparable model scales, obtaining +5.5 gIoU on the ReasonSeg validation set and +4.5 cIoU on the test set.

2 Related Work

2.1 Multimodal RLVR

Recent attempts to apply RLVR to MLLMs have primarily focused on tasks where verifiable rewards can be readily constructed and where advanced reasoning ability is required.

Early studies applied GRPO to general multimodal reasoning tasks such as VQA, using ground-truth answers as verifiable rewards (e.g., VisualThinker-R1-Zero [40], Vision-R1 [9]). Inspired by the success of RLVR in mathematical reasoning for LLMs, subsequent works validated its effectiveness in multimodal mathematical reasoning tasks [4, 22].

Beyond general reasoning, RLVR has also been explored in more vision-centric multimodal tasks. For instance, R1-V [2] and CrowdVLM-R1 [33] applied RLVR to visual counting, while Visual-RFT [20] and VLM-R1 [29] introduced RLVR to Referring Expression Comprehension (REC). In these works, geometrically verifiable metrics such as IoU or distance-based measures were employed as rewards, demonstrating that RLVR can effectively improve geometric accuracy. RLVR has also been extended to domain-specific or non-visual modalities, including medical tasks (MedVLM-R1 [23], ChestX-Reasoner [5]), GUI interaction (UI-R1 [21], InfiGUI-R1 [17]), video understanding (VideoChat-R1 [13]), and audio tasks (R1-AQA [12]), showing its capability to elicit structured reasoning and world knowledge across diverse domains.

However, most prior studies construct task-specific reward functions and directly apply GRPO, with limited exploration of improvements at the objective level. As a result, they inherit known issues of GRPO, such as training instability and entropy collapse, and risk converging to suboptimal local solutions due to a fixed objective design.

In this work, we introduce a multimodal RLVR framework into reasoning segmentation, a task that demands both visual precision and complex reasoning. Furthermore, by investigating objective-level improvements inspired by advances in language-based RLVR, we aim to provide insights into improving learning stability and efficiency not only for reasoning segmentation but also for broader multimodal RLVR applications.

2.2 Reasoning Segmentation

LISA [10] introduced a model architecture that integrates an MLLM with a mask decoder by leveraging features associated with the generation of a special `<SEG>` token in the LLM, enabling reasoning segmentation. FoodLMM [36] extended this structure and introduced a multitask learning framework that jointly trains reasoning segmentation alongside tasks such as question answering, recipe generation, and nutrition estimation in the food domain, achieving complementary performance improvements. However, as these models are primarily trained via Supervised Fine-Tuning (SFT), they tend to struggle particularly in OOD reasoning scenarios.

To address the limitations of SFT, multimodal RLVR has been introduced into reasoning segmentation. Seg-Zero [18] explicitly incorporates reasoning traces through format rewards and constructs geometric rewards based on bounding boxes and points prediction accuracy, applying GRPO-based reinforcement learning to achieve strong OOD generalization. LENS [41] proposed an end-to-end architecture that integrates MLLMs and mask decoders at the feature level and demonstrated the effectiveness of combining SFT with GRPO using geometric losses and reward-based optimization.

Nevertheless, the reinforcement learning strategies explored in these works remain limited to GRPO, leaving challenges such as entropy collapse, reduced exploration in reasoning, and training instability unresolved. In this study, we introduce objective-level improvements to GRPO to enhance learning stability and efficiency. Moreover, we incorporate mechanisms that explicitly enhance perceptual capability alongside reasoning performance, proposing an effective RL-based reasoning segmentation framework.

3 Method

This section describes our reasoning segmentation framework PARS: Perception-Aware Reasoning Segmentation. In Section 3.1, we present the model architecture used in this framework. Then, in Section 3.2, we introduce a new group-based policy optimization method, Adaptive Perception-Aware Policy Optimization (AP2O). Finally, in Section 3.3, we describe the reward design, including a non-repetition reward, a format reward, and an accuracy reward that supports multiple objects.

3.1 Model Architecture

Our reasoning segmentation model follows the decoupled architecture of Seg-Zero [18]. Specifically, we use an MLLM as the reasoning module and a promptable segmentation model as the segmentation module, and integrate them at the token level (Fig. 2).

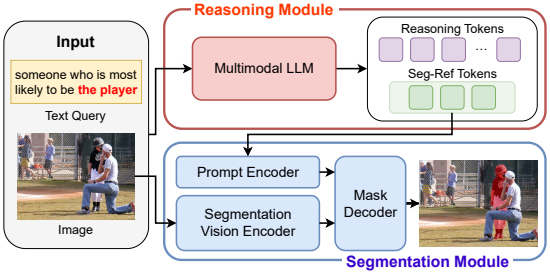


Fig. 2: Overview of our model architecture.

The MLLM policy π_θ takes an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a text query q. Within π_θ, a vision encoder and a vision-language connector first produce a sequence of visual token embeddings from I, which are then concatenated with the token embeddings of q and fed into the LLM. The response token sequence $S = [s_1, \ldots, s_T]$ is generated autoregressively. The response S contains reasoning tokens S_{reason} enclosed by `<think>` and segmentation-reference tokens S_{segref}

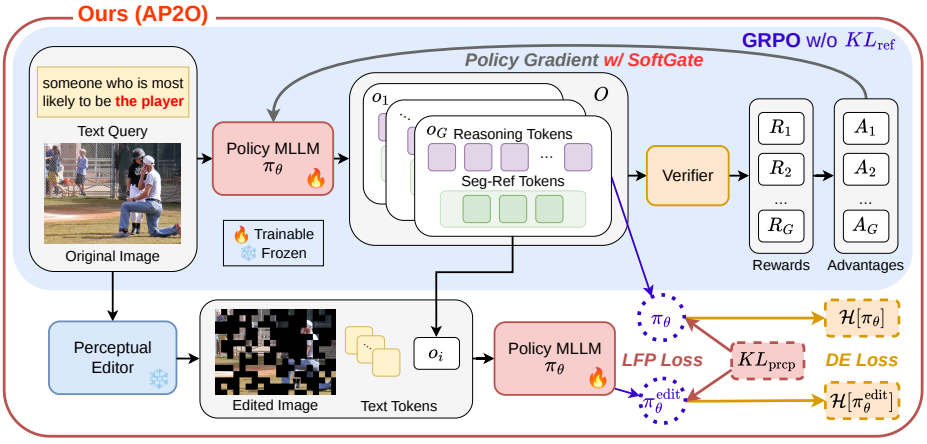


Fig. 3: Overview of our training framework.

enclosed by `<answer>`. This format is reliably acquired through the prompting strategy and the format reward described in Section 3.3.

Next, the segmentation-reference tokens S_{segref} are parsed into a set of segmentation references consisting of bounding boxes and representative points: $\hat{\mathcal{P}} = \{(\hat{b}_i, \hat{p}_i)\}_{i=1}^{K}$. Then, $\hat{\mathcal{P}}$ together with the image I is fed into the segmentation module $\mathcal{S}$. Inside $\mathcal{S}$, a prompt encoder extracts positional features for the boxes $\hat{b}_i$ and points $\hat{p}_i$ after applying positional embeddings. A hierarchical ViT-based vision encoder extracts multi-scale visual features from I. Finally, a decoder-style Transformer with bidirectional cross-attention fuses positional and visual features, and a deconvolution layer produces a binary mask $M \in \{0,1\}^{H \times W}$.

During training, rewards are computed only from the MLLM response S and the parsed segmentation references $\hat{\mathcal{P}}$. Therefore, predicting the final segmentation mask with $\mathcal{S}$ is not required during training and this inference path is activated only at evaluation time.

3.2 Adaptive Perception-Aware Policy Optimization (AP2O)

To jointly enhance reasoning and perception, both of which are crucial for reasoning segmentation, we propose a new group-based policy optimization algorithm AP2O: Adaptive Perception-Aware Policy Optimization (Fig. 3).

Soft-gated surrogate objective To promote improvements in reasoning performance, we are the first to introduce a SAPO-style soft-gated surrogate objective [6] into reasoning segmentation. We further combine it with dynamic sampling [39] to prevent vanishing effective gradients, remove the KL regularization term to avoid excessive regularization, and introduce a new perception loss to encourage attention to visual evidence and accurate localization, thereby improving visual grounding.

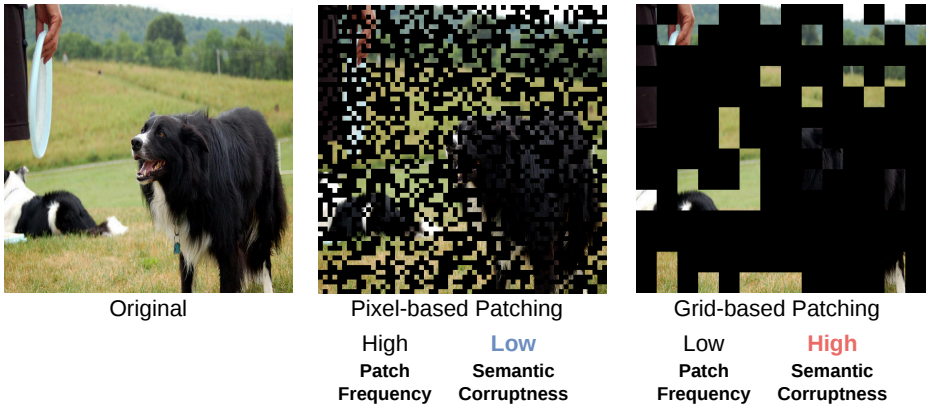


Fig. 4: Construction of low-frequency patches and comparison of semantic corruption.

Our objective is defined as:

$$\mathcal{J}_{\text{AP2O}}(\theta) = \mathbb{E}_{(q,I)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q,I)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg\{ f_{i,t}(r_{i,t}(\theta))\widehat{A}_{i,t} + \gamma\,\mathbb{D}_{\text{KL}}\big[\pi_\theta \parallel \pi_\theta^{\text{edit}}\big] - \eta_1\,\mathcal{H}[\pi_\theta] - \eta_2\,\mathcal{H}\big[\pi_\theta^{\text{edit}}\big]\Bigg\} \tag{1}$$

where the soft-gating function $f_{i,t}(x)$ is defined as:

$$f_{i,t}(x) = \frac{1}{1+e^{-\tau_{i,t}\,(x-1)}}\cdot\frac{4}{\tau_{i,t}}, \quad \tau_{i,t} = \begin{cases}\tau_{\text{pos}}, & \text{if } \widehat{A}_{i,t} > 0\\ \tau_{\text{neg}}, & \text{otherwise}\end{cases} \tag{2}$$

τ_{pos} and τ_{neg} are the temperature parameters for positive and negative tokens in $f_{i,t}(x)$.

The KL term $\mathbb{D}_{\text{KL}}[\pi_\theta||\pi_\theta^{\text{edit}}]$ is a perception loss, which is computed using the low-variance unbiased estimator of KL divergence [27]:

$$\mathbb{D}_{\text{KL}}[\pi_\theta||\pi_\theta^{\text{edit}}] = r_i^{\text{prcp}}(\theta) - \log r_i^{\text{prcp}}(\theta) - 1, \tag{3}$$

where

$$r^{\text{prcp}}(\theta) = \frac{\pi_\theta(o \mid q, I)}{\pi_\theta(o \mid q, I^{\text{edit}})}. \tag{4}$$

Moreover, $\mathcal{H}[\pi_\theta] = \log \pi_\theta(o|q,I)$ and $\mathcal{H}[\pi_\theta^{\text{edit}}] = \log \pi_\theta(o|q,I^{\text{edit}})$ correspond to Double Entropy Loss [32], which we refer to as the Original Entropy (OE) loss and Masked Entropy (ME) loss, respectively. γ, η_1, η_2 are hyperparameters controlling the weights of each term.

Low-Frequency Patch Perception Loss (LFP Loss) To encourage attention to visual evidence and accurate localization, we introduce a new perception objective, the Low-Frequency Patch Perception Loss (LFP Loss).

PAPO reports that Implicit Perception Loss (IP Loss) based on high-frequency perturbations, such as Gaussian noise, fails to sufficiently induce semantic corruption and are therefore inadequate for strengthening attention to visual evidence [32]. Indeed, object shape and global layout are known to be primarily represented by low-frequency components [30], and high-frequency noise may fail to remove semantic information. However, IP Loss typically fixes the patch size in absolute pixel units, which makes the mask pattern effectively higher-frequency when image resolution is large.

To preserve low-frequency corruption regardless of image resolution, we set the patch size relative to the image size. Specifically, we define the side length l_p of each patch $p_{u,v}$ on a $u \times v$ grid as:

$$l_p = \frac{\min(H, W)}{\alpha}, \tag{5}$$

where H and W are the image height and width, and α is a hyperparameter representing the number of grid units placed along the shorter side. Using l_p and a random masking operator $\mathcal{M}$, we obtain the edited image as $I^{\text{edit}} = \mathcal{M}(I, l_p)$.

Fig. 4 illustrates examples of patching with a fixed pixel size and with grid-based relative patch sizing. With fixed-size patching (middle), the dog and frisbee remain barely recognizable, indicating that object identity and global layout are not sufficiently disrupted. Consequently, high-frequency masking based on fixed pixel lengths results in limited semantic corruption. In contrast, grid-based masking (right) makes it difficult to perceive that the image contains two dogs and a frisbee or to identify their approximate locations, demonstrating that low-frequency masking induced by resolution-aware patch sizing yields stronger semantic corruption.

By introducing such low-frequency masking and increasing the rate at which semantically meaningful visual evidence is obscured, our approach encourages the policy to attend more strongly to semantically important visual cues.

3.3 Reward Design

Following the reward designs of Seg-Zero and VisionReasoner [19], we design a unified reward that simultaneously encourages (i) adherence to the required output format, (ii) geometric correctness of multiple segmentation references (bounding boxes and representative points), and (iii) suppression of redundant repetition.

The final reward R is defined as a weighted sum of the format reward R_{fmt}, the accuracy reward R_{acc}, and the non-repetition reward R_{rep}:

$$R = \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{acc}} R_{\text{acc}} + \lambda_{\text{rep}} R_{\text{rep}}. \tag{6}$$

Here, $\lambda_{\text{fmt}}, \lambda_{\text{acc}}, \lambda_{\text{rep}}$ control the trade-off between format compliance, geometric accuracy, and repetition avoidance.

Format reward R_{fmt} The format reward encourages the response to follow the required structured output format. Specifically, we evaluate whether the response satisfies the global response structure consisting of reasoning and answer segments, and the validity of the segmentation references decoded from the answer segment. The format reward is defined as a weighted sum of the token-level format reward R_{tf} and the segmentation-reference format reward R_{sf}:

$$R_{\text{fmt}} = \lambda_{\text{tf}} R_{\text{tf}} + \lambda_{\text{sf}} R_{\text{sf}}. \tag{7}$$

Token-level format reward R_{tf} R_{tf} evaluates whether the response S satisfies a strictly separated format consisting of a reasoning part followed by an answer part, and whether they appear in the required order. It returns 1 if the condition holds and 0 otherwise. In practice, we verify the full response format using a regular expression:

$$R_{\text{tf}} = \mathbb{I}[\texttt{full_match}(S, \rho_{\text{tf}})]\,. \tag{8}$$

where

$$\rho_{\text{tf}} = \texttt{<think>.*?</think>\textbackslash s*<answer>.*?</answer>}$$

and $\texttt{full_match}(S, \rho_{\text{tf}})$ is a predicate that checks whether the response S matches the regular expression ρ_{tf} as a whole. ρ_{tf} corresponds to the response format in which `<think>...</think>` is immediately followed by `<answer>...</answer>`.

Segmentation-reference format reward R_{sf} We parse `<answer>...</answer>` as JSON and encourage each decoded reference to satisfy $\hat{b}_i \in \mathbb{R}^4$ and $\hat{p}_i \in \mathbb{R}^2$. If parsing fails, the reward is set to 0:

$$R_{\text{sf}} = \frac{1}{K} \sum_{i=1}^{K} \Big(\mathbb{I}[\hat{b}_i \in \mathbb{R}^4] + \mathbb{I}[\hat{p}_i \in \mathbb{R}^2] \Big). \tag{9}$$

Accuracy reward R_{acc} via matching multiple objects The accuracy reward assumes multiple object references and evaluates geometric consistency after matching predicted references $\hat{\mathcal{P}}$ to ground-truth references $\mathcal{G}$. We apply thresholding on IoU, L1 distance for bounding boxes, and L2 distance for representative points:

$$R_{\text{acc}} = \lambda_{\text{iou}} R_{\text{IoU}}^{\text{bbox}} + \lambda_{\text{l1}} R_{\text{L1}}^{\text{bbox}} + \lambda_{\text{pt}} R_{\text{L2}}^{\text{pt}}. \tag{10}$$

Pairwise criteria For each pair of predicted and ground-truth references (i, j), we define the following binary rewards:

$$r_{ij}^{\text{iou}} = \mathbb{I}\Big[\text{IoU}(\hat{b}_i, b_j) > \tau_{\text{IoU}}\Big], \tag{11}$$

$$r_{ij}^{\text{l1}} = \mathbb{I}\Big[\|\hat{b}_i - b_j\|_1 < \epsilon_{\text{bbox}}\Big], \tag{12}$$

$$r_{ij}^{\text{pt}} = \mathbb{I}\Big[\|\hat{p}_i - p_j\|_2 < \epsilon_{\text{pt}} \ \wedge \ \hat{p}_i \in \hat{b}_i\Big]. \tag{13}$$

For r_{ij}^{pt}, we only allow positive reward when the representative point lies inside the predicted box $\hat{b}_i$, further improving geometric consistency.

Hungarian matching We match K predictions to N ground truths via the Hungarian algorithm. Specifically, we seek an assignment that maximizes

$$s_{ij} = r_{ij}^{\text{iou}} + r_{ij}^{\text{l1}} + r_{ij}^{\text{pt}}, \tag{14}$$

equivalently minimizing the cost $3 - s_{ij}$. Let $\mathcal{M} \subseteq \{1, \ldots, K\} \times \{1, \ldots, N\}$ denote the resulting matched set. We then define

$$R_{\text{IoU}}^{\text{bbox}} = \frac{1}{\max(N, K)} \sum_{(i,j) \in \mathcal{M}} r_{ij}^{\text{iou}}, \tag{15}$$

$$R_{\text{L1}}^{\text{bbox}} = \frac{1}{\max(N, K)} \sum_{(i,j) \in \mathcal{M}} r_{ij}^{\text{l1}}, \tag{16}$$

$$R_{\text{L2}}^{\text{pt}} = \frac{1}{\max(N, K)} \sum_{(i,j) \in \mathcal{M}} r_{ij}^{\text{pt}}. \tag{17}$$

This normalization stabilizes the reward scale even when the number of predictions K differs from the number of ground-truth objects N.

Non-repetition reward R_{rep} The non-repetition reward R_{rep} is introduced to suppress excessive repetition of the same sentence within a response. We split the response into sentences by periods to obtain a set of unique sentences $\mathcal{U}$. Let $c(u)$ be the occurrence count of sentence $u \in \mathcal{U}$. We penalize responses whose maximum duplication count exceeds a threshold D_{dup}:

$$R_{\text{rep}} = \mathbb{I}\big[\max_{u \in \mathcal{U}} c(u) \leq D_{\text{dup}}\big]. \tag{18}$$

4 Experiments

4.1 Experimental Setup

In our experiments, we use Qwen2.5-VL 3B [1] and SAM2 (Hiera-L) [25] as the reasoning and segmentation modules, respectively. This choice ensures fair comparison with Seg-Zero while keeping the training setup within practical memory constraints. During training, only the MLLM is trainable, while all other weights are frozen.

For the soft-gating function, we set the temperature parameters for positive and negative tokens to $\tau_{\text{pos}} = 1.0$ and $\tau_{\text{neg}} = 1.05$, respectively. For computing R_{acc}, we set the thresholds to $\tau_{\text{IoU}} = 0.5$, $\epsilon_{\text{bbox}} = 10$, and $\epsilon_{\text{pt}} = 30$. We set $D_{\text{dup}} = 2$ for R_{rep}. Following prior work [18, 19], we set all coefficients for the reward terms to 1 to match their configurations. We use AdamW for optimization with a learning rate of 1.0×10^{-6}. The update batch size is set to 32, and training is performed on 8 A6000 GPUs.

4.2 Training Dataset

We use the region-reference annotations provided in VisionReasoner [19], which are built from multiple segmentation datasets including LVIS [8], RefCOCOg [38], gRefCOCO [16], and LISA++ [35].

Each training sample consists of an image, a referring expression, and multiple region references represented by a bounding box and a representative point per object. These region references are derived from instance-level masks by computing the enclosing bounding box and the mask centroid. We use approximately 7k training samples.

4.3 Evaluation Benchmarks

We primarily evaluate on ReasonSeg [10], using its validation and test sets as standard benchmarks for reasoning segmentation. Note that we do not use any ReasonSeg data during training, and thus our evaluation is conducted in a zero-shot manner. This setup tests robustness under OOD data.

4.4 Evaluation Metrics

Following prior reasoning segmentation works, we report gIoU and cIoU. Given predicted masks $\{\hat{M}_i\}_{i=1}^N$ and ground-truth masks $\{M_i\}_{i=1}^N$ over N images, the metrics are defined as

$$\text{gIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{M}_i \cap M_i|}{|\hat{M}_i \cup M_i|}, \qquad \text{cIoU} = \frac{\sum_{i=1}^{N}|\hat{M}_i \cap M_i|}{\sum_{i=1}^{N}|\hat{M}_i \cup M_i|}. \tag{19}$$

gIoU averages IoU per image, while cIoU aggregates intersections and unions across the dataset and thus places more weight on large objects. As in prior work, we use gIoU as the primary metric and report cIoU for completeness.

4.5 Results on Reasoning Segmentation

Table 1 summarizes the reasoning segmentation results on ReasonSeg. We compare our method against open-vocabulary segmentation baselines that do not use an MLLM, including OVSeg [14], ReLA [16], X-Decoder [42], SEEM [43], and Grounded-SAM [26]. We also compare against prior methods that use MLLMs, including LISA [10], GLaMM [24], SAM4MLLM [3], SegLLM [31], HRSeg [15], HyperSeg [34], Seg-Zero [18], and Seg-R1 [37].

Methods without MLLMs have limited capability to identify targets based on language conditions, and their gIoU remains in the 20% range. In contrast, SFT-based MLLM methods achieve much higher performance, reaching the high-50% gIoU range (e.g., HRSeg and HyperSeg on val). With reinforcement learning, our method improves substantially over the Qwen2.5-VL 3B + SAM2 baseline: validation gIoU increases from 53.8 to 63.7 (+9.9), and test gIoU increases from 47.6 to 60.0 (+12.4). Compared with existing RL-based methods, our validation

Table 1: Reasoning segmentation results on the ReasonSeg benchmark.

Category	Method	ReasonSeg val		ReasonSeg test	
		gIoU ↑	cIoU ↑	gIoU ↑	cIoU ↑
w/o MLLM	OVSeg	28.5	18.6	26.1	20.8
	ReLA	22.4	19.9	21.3	22.0
	X-Decoder	22.6	17.9	21.7	16.3
	SEEM	25.5	21.2	24.3	18.7
	Grounded-SAM	26.0	14.5	21.3	16.4
w/ MLLM SFT	LISA-7B [CVPR24]	44.4	46.0	36.8	34.1
	LISA-13B [CVPR24]	48.9	46.9	44.8	45.8
	GLaMM 7B [CVPR24]	47.4	47.2	-	-
	SAM4MLLM 7B [ECCV24]	46.7	48.1	-	-
	SegLLM 7B [ICLR25]	57.2	54.3	52.4	48.4
	HRSeg 7B [ACMMM25]	56.7	55.4	54.6	<u>51.2</u>
	HyperSeg 3B [CVPR25]	59.2	**56.7**	-	-
w/ MLLM RL	Qwen2.5VL 3B+SAM2 (w/o RL)	53.8	44.1	47.6	37.4
	Seg-Zero 3B [arXiv25]	58.2	53.1	<u>56.1</u>	48.6
	Seg-R1 3B [NeurIPSW25]	<u>60.8</u>	56.2	55.3	46.6
	PARS 3B (Ours)	**63.7**	<u>56.3</u>	**60.0**	**53.1**

gIoU surpasses Seg-R1 (60.8), and our test cIoU also exceeds HRSeg (51.2), reaching 53.1. These results demonstrate that the proposed training framework improves reasoning segmentation accuracy beyond prior approaches.

We further present qualitative outputs on ReasonSeg test from our model and the GRPO baseline in Fig. 5. Our method successfully identifies the correct target through reasoning and grounds it in the image. The GRPO baseline predicts an easily confusable but incorrect object, while our method correctly identifies the truly implied target region and mentions the appropriate target succinctly. Furthermore, our method successfully estimates the target region even when the ground-truth region is extremely fine-grained. This can be attributed to improved visual recognition capability induced by the perception loss, compared with GRPO. In particular, the GRPO baseline fails to perceive the visual evidence corresponding to the correct region and produces a confused rationale, whereas our model succinctly describes the perceived visual evidence, indicating improved perception.

Overall, these qualitative results suggest that our method improves both perception and reasoning, leading to higher final reasoning segmentation accuracy. Moreover, our method tends to generate fewer tokens at inference time than prior approaches, suggesting that it encourages more efficient reasoning and can reduce computational cost associated with the inference-time token budget.

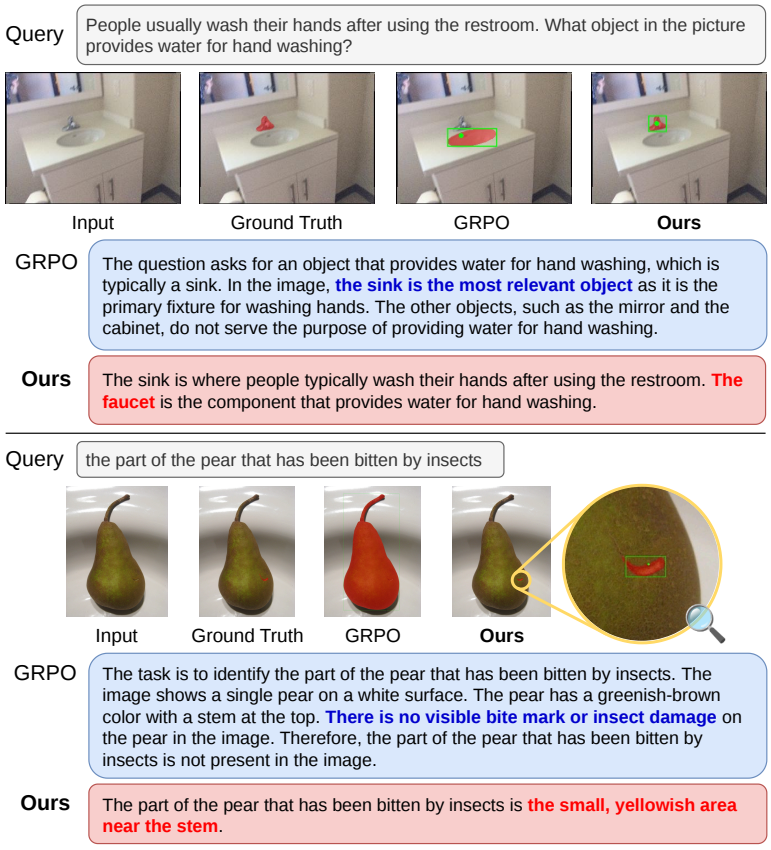


Fig. 5: Qualitative examples on ReasonSeg.

4.6 Effects of Soft Gating and Perception Loss

Table 2 presents the contribution of each component. Using GRPO (Avg. 55.6) as a baseline, introducing only SoftGate (SG) improves the average metric to 57.0, with a notable gain in validation cIoU from 50.0 to 58.5. This suggests that importance-ratio control via soft gating improves the learnability of region estimation accuracy compared with hard clipping. However, test cIoU decreases to 50.7, indicating that further investigation is needed regarding generalization.

When combining Implicit Perception Loss (IP) and Double Entropy Loss (DE) (+IP+DE), we observe improvements on the test set, but when additionally combined with SoftGate (SG+IP+DE), the average metric tends to decrease. In contrast, replacing IP with our proposed Low-Frequency Patch Perception Loss (LFP) in SG+LFP+DE yields the best performance, achieving 63.7 val gIoU and 60.0 test gIoU with the highest average metric of 58.3. These results indicate that stabilized optimization via importance-ratio control and effective perceptual enhancement via an appropriate perception loss complement each other.

Table 2: Impact of key design components.

Setting	Modification			ReasonSeg val		ReasonSeg test		Avg. ↑
	SG	Prcp.	DE	gIoU ↑	cIoU ↑	gIoU ↑	cIoU ↑	
Naive GRPO	✗	✗	✗	60.7	50.0	57.2	54.6	55.6
+IP+DE	✗	IP	✓	59.4	52.2	58.3	55.0	56.2
+SG	✓	✗	✗	61.7	58.5	57.3	50.7	57.0
SG+IP+DE	✓	IP	✓	61.8	52.4	59.0	50.4	55.9
SG+LFP+DE	✓	LFP	✓	63.7	56.3	60.0	53.1	58.3

Table 3: Effect of Double Entropy Loss.

Setting	Modif.		ReasonSeg val		ReasonSeg test		Avg. ↑
	ME	OE	gIoU ↑	cIoU ↑	gIoU ↑	cIoU ↑	
Naive GRPO	✗	✗	60.7	50.0	57.2	54.6	55.6
+ IP + ME Loss	✓	✗	60.0	43.7	58.8	52.7	53.8
+ OE Loss	✓	✓	59.4	52.2	58.3	55.0	56.2

4.7 Effect of Double Entropy Loss

Table 3 analyzes the effect of Double Entropy Loss. Compared with GRPO, adding only Masked Entropy Loss (ME) regularization (+IP+ME Loss) reduces the average performance to 53.8. In contrast, adding Original Entropy Loss (OE) together with ME (+OE Loss) recovers and improves the average performance to 56.2. This suggests that constraining both distributions jointly is most effective for improving region estimation accuracy.

5 Conclusion

In this work, we introduced PARS, a unified framework that integrates perceptual enhancement into RL-based reasoning segmentation. Within this framework, we proposed AP2O, a novel group-based policy optimization method, and demonstrated that it improves training efficiency while enhancing perceptual capability for reasoning segmentation.

Future work includes exploring reward designs that explicitly account for the content of the reasoning process and directly incorporate pixel-level segmentation masks, as these aspects were beyond the primary focus of this study. More broadly, tighter integration of objective functions and reward design may lead to more effective training strategies for reasoning segmentation. In addition, a deeper analysis of the characteristics of different perception losses could provide further insights and lead to additional improvements in perceptual performance.

Acknowledgements This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923 (2025) 10
2. Chen, L., Li, L., Zhao, H., Song, Y., Vinci: R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V (2025), accessed: 2025-02-02 4
3. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: Proc. of European Conference on Computer Vision (2024) 11
4. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In: Advances in Neural Information Processing Systems (2025) 4
5. Fan, Z., Liang, C., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. arXiv preprint arXiv:2504.20930 (2025) 4
6. Gao, C., Zheng, C., Chen, X.H., Dang, K., Liu, S., Yu, B., Yang, A., Bai, S., Zhou, J., Lin, J.: Soft adaptive policy optimization. arXiv preprint arXiv:2511.20347 (2025) 6
7. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 2
8. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 5356–5364 (2019) 11
9. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025) 4
10. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning segmentation via large language model. In: Proc. of IEEE Computer Vision and Pattern Recognition (2024) 1, 4, 11
11. Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L.J.V., Liu, A., Dziri, N., Lyu, S., et al.: Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124 (2024) 2
12. Li, G., Liu, J., Dinkel, H., Niu, Y., Zhang, J., Luan, J.: Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. arXiv preprint arXiv:2503.11197 (2025) 4
13. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025) 4
14. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 7061–7070 (2023) 11
15. Lin, W., Ma, Y., Sun, X., He, S., Ji, J., Cao, L., Ji, R.: Hrseg: High-resolution visual perception and enhancement for reasoning segmentation. In: Proc. of ACM International Conference Multimedia. pp. 3202–3211 (2025) 11

16. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 23592–23601 (2023) 11
17. Liu, Y., Li, P., Xie, C., Hu, X., Han, X., Zhang, S., Yang, H., Wu, F.: Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. arXiv preprint arXiv:2504.14239 (2025) 4
18. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025) 2, 5, 10, 11
19. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: VisionReasoner: Unified visual perception and reasoning via reinforcement learning. arXiv preprint arXiv:2505.12081 (2025) 8, 10, 11
20. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-RFT: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025) 4
21. Lu, Z., Chai, Y., Guo, Y., Yin, X., Liu, L., Wang, H., Xiao, H., Ren, S., Xiong, G., Li, H.: Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. arXiv preprint arXiv:2503.21620 (2025) 4
22. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025) 4
23. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347 (2025) 4
24. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: GLaMM: Pixel grounding large multimodal model. In: Proc. of IEEE Computer Vision and Pattern Recognition (2024) 11
25. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 10
26. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) 11
27. Schulman, J.: Approximating kl divergence. http://joschu.net/blog/kl-approx.html (2020), accessed: 2025-12-27 7
28. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., Guo, D.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 2
29. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., Xu, R., Zhao, T.: VLM-R1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025) 4
30. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 8684–8694 (2020) 8
31. Wang, X., Zhang, S., Li, S., Kallidromitis, K., Li, K., Kato, Y., Kozuka, K., Darrell, T.: Segllm: Multi-round reasoning segmentation. arXiv preprint arXiv:2410.18923 (2024) 11

32. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., et al.: Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448 (2025) 7, 8
33. Wang, Z., Feng, P., Lin, Y., Cai, S., Bian, Z., Yan, J., Zhu, X.: CrowdVLM-R1: Expanding r1 ability to vision language model for crowd counting using fuzzy group relative policy reward. arXiv preprint arXiv:2504.03724 (2025) 4
34. Wei, C., Zhong, Y., Tan, H., Liu, Y., Zhao, Z., Hu, J., Yang, Y.: Hyperseg: Towards universal visual segmentation with large language model. arXiv preprint arXiv:2411.17606 (2024) 11
35. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: Lisa++: An improved baseline for reasoning segmentation with large language model. arXiv preprint arXiv:2312.17240 (2023) 11
36. Yin, Y., Qi, H., Zhu, B., Chen, J., Jiang, Y.G., Ngo, C.W.: FoodLMM: A versatile food assistant using large multi-modal model. arXiv preprint arXiv:2312.14991 (2024) 4
37. You, Z., Wu, Z.: Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. arXiv preprint arXiv:2506.22624 (2025) 11
38. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Proc. of European Conference on Computer Vision. pp. 69–85 (2016) 11
39. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W., Zhang, Y., Yan, L., Qiao, M., Wu, Y., Wang, M.: DAPO: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025) 6
40. Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., Hsieh, C.J.: R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132 (2025) 4
41. Zhu, L., Ouyang, B., Zhang, Y., Cheng, T., Hu, R., Shen, H., Ran, L., Chen, X., Yu, L., Liu, W., et al.: Lens: Learning to segment anything with unified reinforced reasoning. arXiv preprint arXiv:2508.14153 (2025) 5
42. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 15116–15127 (2023) 11
43. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36**, 19769–19782 (2023) 11