Actor as Its Own Critic: Unifying Region Understanding and Localization via CycleGRPO

Xin Zhang[1*], Haochen Wang[2*], Yikang Zhou[4], Zhuochen Wang[2], Xiangtai Li[3], and Robby T. Tan[1]

[1] National University of Singapore
[2] University of Chinese Academy of Sciences
[3] Nanyang Technological University
[4] Wuhan University

Abstract. This paper introduces Actor as Its Own Critic, a unified reinforcement learning framework, Cycle Group Relative Policy Optimization (CycleGRPO), that *jointly* optimizes region understanding and localization for Multimodal Large Language Models (MLLMs). Unlike existing separate pipelines, *we leverage the inherent duality between the two tasks to construct a self-evaluating reinforcement learning paradigm: "region → text → region"*. Specifically, a single MLLM first acts as the actor to generate region captions, then immediately transitions to a critic to ground its generated text back in the spatial domain. Therefore, CycleGRPO requires only region inputs, *e.g.*, masks or bounding boxes, entirely bypassing the need for textual ground truths. A quality-aware token-level cycle-consistency reward is employed to assess the semantic discriminability of text captions via their physical localization accuracy. Empirically, built upon SAMTok, our CycleGRPO framework successfully bootstraps both capabilities simultaneously. Without any task-specific fine-tuning, the framework yields consistent performance gains across a wide range of benchmarks, including region captioning, region VQA, grounded dialogue, and referring segmentation. Overall, CycleGRPO offers a straightforward and scalable way to advance pixel-level capabilities in MLLMs. Code and models are released at https://github.com/devinxzhang/CycleGRPO.

Keywords: Multimodal Large Language Models · Reinforcement Learning · Region Understanding · Referring Localization

1 Introduction

MLLMs [2, 19, 38, 42] have shown strong ability in general visual-language understanding. Yet they still lack fine-grained capabilities required for real-world perception. Among these, region understanding [7, 16, 22, 31, 46, 50, 57] and referring localization [9, 14, 18, 45, 53–55, 57] serve as two foundational abilities. The former aims to describe and comprehend given regions, such as their appearance,

* Equal contribution. Project Lead: Xiangtai Li

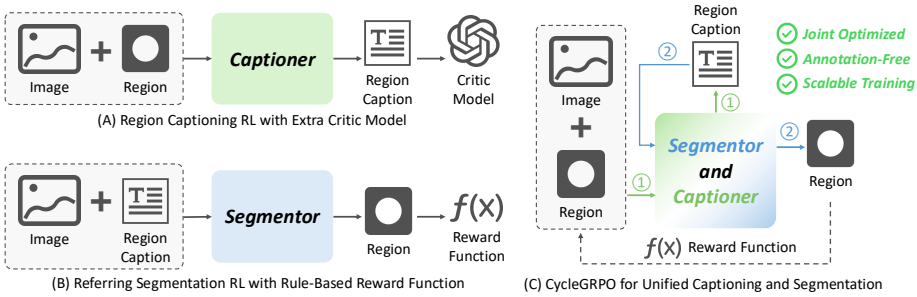

Fig. 1: Concept comparison of Reinforcement Learning paradigms for region-level tasks. (A) Captioning tasks typically rely on external LLMs to judge the text quality, which is vulnerable to reward hacking. **(B)** Localization tasks use objective rule-based functions (*e.g.*, IoU). Both (A) and (B) heavily depend on *unscalable* high-quality text-mask pairs. **(C)** We let a single model act as both the actor (generating captions) and its own critic (referring its own text back into masks). This cycle consistency provides intrinsic reward signals, bypassing external judges and the need for text ground truths.

structure, and spatial relation with others. The latter (including grounding and segmentation) seeks to locate or segment the target region based on free-form language queries. Together, they build the core of fine-grained visual reasoning and are critical for various downstream applications.

Conventional approaches usually optimize these two tasks *separately*, as demonstrated in Figure 1 (A) and (B). We observe that region understanding and referring localization are essentially *dual problems*. Their symmetric nature naturally forms a closed reconstruction pipeline: "region → text → region". Crucially, *a genuinely discriminative caption must contain sufficient unique details to seamlessly guide the model back to the exact initial mask*. Therefore, by exploiting this cycle consistency, the quality of text generation can be evaluated intrinsically using spatial localization accuracy. This elegant formulation simultaneously *obsoletes the need for expensive human annotations*, low-quality auto-generated data, and external LLM judges, unlocking a self-evaluating and highly scalable learning paradigm, as illustrated in Figure 1 (C).

Motivated by this insight, we present Cycle Group Relative Policy Optimization (**CycleGRPO**), a scalable reinforcement learning (RL) pipeline that fundamentally instantiates this duality. Built upon the discrete mask representation of SAMToK [57], which makes fine-grained RL for segmentation feasible, CycleGRPO *incentivizes* the MLLM to produce discriminative captions via cycle-consistent rewards. At its core, CycleGRPO embodies an "Actor as Its Own Critic" paradigm, eliminating the conventional reliance on disparate reward models or task-specific heads. Specifically, the model first acts as the *Actor* to generate a set of candidate region descriptions. Then, the same model transitions into its own *Critic*, which executes a rollout to localize the generated text back into the spatial domain. By explicitly optimizing this intrinsic cycle consistency,

e.g., the reconstructive Intersection-over-Union (IoU) between input regions and predicted regions, CycleGRPO jointly bootstraps region understanding and localization in a fully autonomous manner.

Empirically, we comprehensively evaluate CycleGRPO across a diverse set of tasks, including region captioning, region VQA, grounded dialogue, and referring segmentation. Our method achieves remarkable performance gains on GRES [18], delivering a 7.0% increase in average gIoU compared to the SAMTok baseline. On GroundingSuite [9], CycleGRPO improves the overall grounding accuracy from 57.5% to 67.6%, with a significant 8.5% boost on the challenging Part split. Furthermore, the framework demonstrates superior reasoning and synchronization capabilities, yielding a 5.8% average improvement on DLC-Bench [16], a 0.9% overall gain on GAR-Bench-VQA [30], and a 6.5 increase in CIDEr on the GCG benchmark [22]. Notably, these consistent improvements are achieved without any task-specific fine-tuning on these datasets, underscoring that our approach provides a straightforward and scalable way to advance pixel-level capability in MLLMs by bypassing the bottleneck of textual annotations. In summary, our key contributions are:

- We propose **CycleGRPO**, a unified reinforcement learning framework that fundamentally instantiates the duality between region understanding and localization, enabling a fully autonomous and scalable training pipeline.
- We introduce an *Actor-as-Own-Critic* paradigm in which a single MLLM self-evaluates its text generation using spatial consistency rewards, thereby eliminating reliance on expensive human annotations or external LLM judges.
- Extensive experiments demonstrate that CycleGRPO successfully bootstraps both capabilities simultaneously, achieving highly competitive performance against heavily supervised baselines on multiple fine-grained benchmarks.

2 Preliminary

GRPO for Region Localization and Captioning. RL has been widely adopted to align MLLMs with human preferences and task-specific metrics. Recently, GRPO [24] has emerged as a highly efficient alternative to traditional Proximal Policy Optimization (PPO). By evaluating the relative quality within a sampled group of outputs, GRPO eliminates the need for a parameterized value network (critic), significantly reducing memory overhead while maintaining stable policy updates.

Formally, let q denote a multi-modal input prompt and π_θ denote the MLLM actor. For a given input q, the policy π_θ samples a group of G candidate outputs $\{y_1, y_2, \ldots, y_G\}$. A reward function $R(q, y)$ evaluates these candidates to assign scalar rewards $\{r_1, r_2, \ldots, r_G\}$. The advantage A_i for each output y_i is then computed by standardizing the rewards within the sampled group as $A_i = (r_i - \mu_r)/\sigma_r$, where μ_r and σ_r are the mean and standard deviation of the group rewards, respectively. The model is optimized by maximizing an objective that incorporates a clipping mechanism and a Kullback-Leibler (KL) divergence

penalty to prevent the policy from deviating excessively from a reference model π_{ref}:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{y_i\}_{i=1}^{G}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Big(\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta(y_i|q) \| \pi_{\text{ref}}(y_i|q)) \Big) \Bigg], \tag{1}$$

where $\rho_i = \frac{\pi_\theta(y_i|q)}{\pi_{\theta_{\text{old}}}(y_i|q)}$ is the probability ratio, ϵ is the clipping threshold, and β controls the KL penalty.

When applying this framework to fine-grained region tasks, region understanding and localization are typically treated as isolated optimization problems (as contrasted in Figure 1 (A) and (B)):

- **Region Captioning**: The input $q_{\text{cap}} = (I, M)$ consists of an image I and a region mask M. The actor generates a text caption $y_{\text{cap}} = C$. The reward r typically relies on external textual ground truths (e.g., CIDEr) or expensive LLM-as-a-judge.
- **Region Localization**: The input $q_{\text{loc}} = (I, C)$ consists of an image and a text description C. The policy generates a spatial mask $y_{\text{loc}} = \hat{M}$. The reward r is often defined by the Intersection over Union (IoU) between $\hat{M}$ and the ground-truth mask M.

While effective, this isolated paradigm imposes a severe bottleneck: it inherently necessitates vast amounts of high-quality, human-annotated text-region pairs for supervision. To overcome this, we seek a unified representation that allows these two tasks to interact within a closed, self-supervised loop.

Unified Mask Token Representation via SAMTok. A significant challenge in applying RL to fine-grained region tasks lies in the representation gap: traditional MLLMs often separate discrete text generation from continuous mask decoding via external segmentation heads. To bridge this gap, we build our framework upon the SAMTok architecture [57].

The core of SAMTok is a discrete mask tokenizer that quantizes an arbitrary spatial mask M into a compact sequence of discrete mask tokens T_M. By injecting these tokens into the LLM's extended vocabulary $\mathcal{V}$, a region mask is effectively translated into a format that the model natively understands. This tokenization reframes region localization as a standard autoregressive next-token prediction task, entirely eliminating the need for task-specific heads. Consequently, the MLLM can uniformly model the probability of generating either a text caption C or a spatial mask M within the *same* probability space:

$$P(y|q) = \prod_{j=1}^{L} \pi(y_j | y_{<j}, q), \quad y_j \in \mathcal{V} \tag{2}$$

where y can represent either a sequence of text tokens or mask tokens. This unified discrete representation serves as the structural prerequisite for our frame-

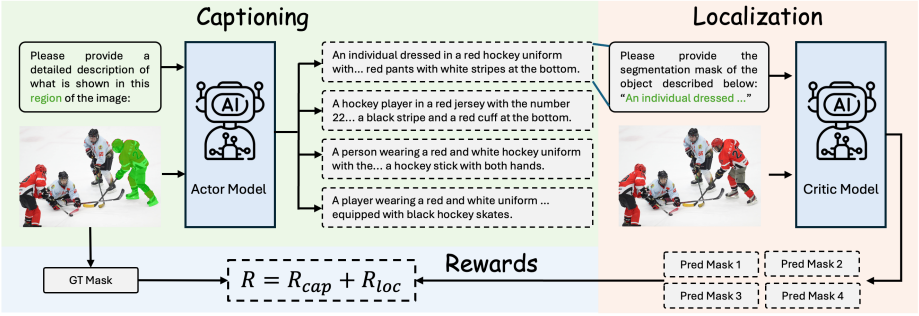

Fig. 2: Overview of the CycleGRPO Pipeline. Our framework leverages a single MLLM to instantiate a self-evaluating closed loop. **Left (Phase 1: Region Captioning):** Acting as the *Actor*, the model takes an image and a target region M to generate a group of candidate text descriptions $\{C_1, \ldots, C_G\}$. **Right (Phase 2: Localization Rollout):** Symmetrically acting as its own *Critic*, the exact same model receives each candidate and attempts to ground it back into the spatial domain, yielding multiple predicted reconstructions $\{\hat{M}_{i,k}\}$. **Bottom:** The semantic quality of the captions is materialized through the spatial consistency between the predicted and original regions. The framework computes a joint cyclic reward based on this alignment, optimizing both understanding and grounding capabilities simultaneously without requiring external text annotations.

work, allowing the same model to process both understanding and localization tasks using a single, consistent vocabulary.

Unified Bounding Box Token Representation. Current advanced open-source MLLMs [1, 2, 38] already support unified token representation for text and bounding boxes. They serve as box coordinates, treated as regular text tokens, eliminating the need for dedicated detection heads. To demonstrate the generalization capabilities of our CycleGRPO, we also conduct experiments on the Qwen3-VL [1] benchmark for region understanding and grounding tasks.

3 CycleGRPO: Actor as Its Own Critic

In this section, we use region *understanding* and *segmentation* with SAMTok [57] by default to illustrate our framework, since unifying region understanding and grounding with Qwen3-VL [1] is nearly the same.

Duality of Region Captioning and Localization. Fundamentally, region captioning (understanding) and region localization (segmentation/grounding) can be formulated as dual problems that perform inverse transformations between visual-spatial and textual-semantic domains. Given an image I, region captioning aims to map a spatial region, defined by a binary mask M, to a discriminative text description C. Conversely, region localization grounds a description C back into the physical space, predicting the corresponding mask $\hat{M}$.

Building upon the unified vocabulary enabled by SAMTok (Sec. 2), we can reformulate these two mappings as conditional probability distributions modeled

by a *single* policy π_θ. The duality implies that these tasks are inherently symmetric and mutually conditional: *a high-quality caption C should contain sufficient discriminative details to enable the precise reconstruction of its original region.*

Formally, if π_θ generates an optimal caption C^* via the understanding mapping $C^* \sim \pi_\theta(\cdot|I, M)$, the inverse mapping $\pi_\theta(\cdot|I, C^*)$ should ideally recover the initial spatial mask. Formally, the dual-task consistency aims to *minimize the expected reconstruction error across the cycle*:

$$\mathcal{L}_{\text{cycle}}(\theta) = \mathbb{E}_{C\sim\pi_\theta(\cdot|I,M)} \left[\mathcal{D}\big(M, \pi_\theta(\cdot|I, C)\big)\right], \tag{3}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes a distance metric (e.g., IoU) in the spatial representation space. The model first acts as an actor to generate a semantic bridge C, and subsequently validates it through spatial reconstruction as a critic.

Unlike traditional RL pipelines that isolate these processes, this dual perspective reveals that the model's localization capability can serve as an *intrinsic*, physically grounded metric for evaluating its own semantic understanding. This insight directly motivates our **Actor-as-Own-Critic** paradigm, as illustrated in Figure 1(C). In this closed loop, the model first acts as the *actor* to generate candidate captions, then immediately transitions to its own *critic* to perform inverse localization. By natively minimizing the cycle discrepancy without external textual ground truth, the framework directly ties the optimization of semantic discriminability to physical grounding accuracy.

The CycleGRPO Framework. Leveraging a unified representation space, we propose **CycleGRPO**. As illustrated in Figure 2, the MLLM π_θ dynamically transitions between a region captioner and a region localizer, forming a self-evaluating closed loop. Given an image I and a target region mask M, the framework executes this cycle via a two-stage rollout process:

Phase 1: Captioning Rollout (Region Understanding). In the first phase, the model performs semantic exploration. We define a *captioning prompt* $q_{\text{cap}} = (I, Q_{\text{cap}}(M))$, where $Q_{\text{cap}}(\cdot)$ is a task-specific instruction template that embeds the target region. The policy π_θ then autoregressively samples a group of G candidate captions:

$$\{C_1, C_2, \ldots, C_G\} \sim \pi_\theta(\cdot \mid q_{\text{cap}}). \tag{4}$$

In the absence of textual ground truths or external judges, these candidates represent a diverse set of semantic hypotheses regarding the visual region.

Phase 2: Localization Rollout (Region Localization). To evaluate these hypotheses, the model transitions into the role of a localizer to close the cycle. For each candidate caption C_i, we construct a *localizing prompt* $q_{\text{loc}}^{(i)} = (I, Q_{\text{loc}}(C_i))$ to query the model for spatial reconstruction.

A key challenge here is the inherent stochasticity of autoregressive decoding: a single greedy or sampled prediction $\hat{M}_i$ may not fully reflect the discriminative potential of C_i. To obtain a more robust proxy for semantic quality, we perform K independent localizing rollouts for each candidate C_i, yielding a set of predicted spatial masks:

$$\{\hat{M}_{i,1}, \ldots, \hat{M}_{i,K}\} \sim \pi_\theta(\cdot \mid q_{\text{loc}}^{(i)}), \quad \forall i \in \{1, \ldots, G\}. \tag{5}$$

By aggregating the spatial accuracy across these K rollouts, the framework obtains a robust estimate of the caption's quality, effectively suppressing the noise inherent in autoregressive sampling.

Through this dual-stage process, the abstract semantic quality of C_i is materialized into concrete spatial reconstructions $\{\hat{M}_{i,k}\}_{k=1}^{K}$. If the generated caption C_i is accurate and unambiguous, the grounding rollout should consistently recover the initial spatial representation M. This materialization allows us to bypass the need for external textual supervision, as the spatial consistency between M and $\hat{M}$ now serves as the foundational signal for the reward mechanism.

The Cycle Reward. The CycleGRPO framework optimizes the model by converting the spatial consistency between the initial region M and its cyclic reconstructions $\{\hat{M}_{i,k}\}$ into training signals for both tasks. For each grounding rollout $\hat{M}_{i,k}$, we compute an alignment score $s_{i,k}$ that measures its spatial consistency with the reference mask M. Following standard practice in visual grounding, we define this score as the Intersection over Union (IoU) between the masks, denoted as $s_{i,k} = \text{IoU}(M, \hat{M}_{i,k})$. Using IoU provides a continuous and physically grounded signal that accurately reflects the quality of the reconstruction. Specifically, this alignment score is then used to supervise both the captioning and grounding phases:

- **Captioning Reward (R_i^{cap}):** The quality of the i-th candidate caption C_i is evaluated by the average IoU of its K corresponding grounding paths: $R_i^{\text{cap}} = \frac{1}{K}\sum_{k=1}^{K} s_{i,k}$. A low R_i^{cap} indicates that the caption is either hallucinated or too vague to facilitate precise reconstruction.
- **Localization Reward ($R_{i,k}^{\text{loc}}$):** For the grounding phase, we weight each trajectory by its parent caption's reward: $R_{i,k}^{\text{loc}} = R_i^{\text{cap}} \cdot s_{i,k}$. This weighting ensures that gradients from high-quality captions are amplified, while those from low-quality or hallucinated text are suppressed.

The final training reward for a complete spatial-semantic cycle is defined as the sum of the two components: $R^{\text{total}} = R^{\text{cap}} + R^{\text{loc}}$. By maximizing this joint reward, the model is incentivized to generate captions that are not only linguistically fluent but also spatially discriminative, while simultaneously refining its grounding precision. This total reward is subsequently standardized within each sampling group to compute the advantages A^{cap} and A^{loc} as required by the GRPO objective (Eq. 1).

4 Experiment

Implementation Details. We implement the CycleGRPO framework using the `verl` RL library. Our base MLLM is SAMTok [57], built upon the Qwen3-VL-4B [1] backbone. Note that we employ the pre-trained version of SAMTok without any further supervised fine-tuning (SFT) or RL on specific downstream datasets. During training, we freeze the vision encoder and fine-tune the projection layer and the LLM parameters. For the training corpus, we sample 20k images with corresponding object masks from DenseWorld [15], supplemented by

Table 1: Results on the mask-to-text task (DLC-Bench [16]). † indicates trained on *in-domain* DAM-1.5M [16] with respect to DLC-Bench [16].

Method	Size	Pos.	Neg.	Avg.
GPT-4o [11]	–	43.4	79.6	61.5
Gemini 2.5 Pro [27]	–	36.5	75.2	55.8
DAM† [16]	3B	52.3	82.2	67.3
SAMTok [57]	4B	43.5	80.4	61.9
+ GRPO	4B	45.0	81.8	63.4
+ CycleGRPO	4B	51.2	84.2	67.7
Δ *v.s.* SAMTok [57]		↑ 7.7	↑ 3.8	↑ 5.8
Δ *v.s.* GRPO		↑ 6.2	↑ 2.4	↑ 4.3

Table 2: Results on text-to-mask task (GroundingSuite [9]). We report the gIoU to evaluate the model's zero-shot generalization capability across diverse spatial grounding scenarios.

Method	Size	Stuff	Part	Multi	Single	All
LISA [12]	7B	85.2	21.2	71.5	42.8	57.6
GLaMM [22]	7B	86.9	16.5	70.4	42.1	57.2
EVF-SAM [56]	1B	85.1	23.1	72.1	54.5	62.6
InstructSeg [41]	3B	56.2	24.2	66.8	51.3	52.5
SAMTok [57]	4B	80.9	12.4	62.0	52.9	57.5
+ GRPO	4B	89.3	16.5	71.6	54.1	62.7
+ CycleGRPO	4B	90.7	20.9	76.3	61.6	67.6
Δ *v.s.* SAMTok [57]		↑ 9.8	↑ 8.5	↑ 14.3	↑ 8.7	↑ 10.1
Δ *v.s.* GRPO		↑ 1.4	↑ 4.4	↑ 4.7	↑ 7.5	↑ 4.9

1k no-target expressions from GRES [18] to enhance discriminative robustness. The model is trained for one epoch with a total batch size of 128. Following the group-sampling strategy of GRPO, we set both the caption group size G and the grounding rollout count K to 6, yielding $G \times K = 36$ spatial trajectories per image to estimate the cycle reward. We use the AdamW optimizer with a learning rate of 1×10^{-6} and a weight decay of 1×10^{-2}.

Dataset and Benchmarks. To evaluate the effectiveness of CycleGRPO, we conduct experiments on two representative task categories: region captioning and region localization. Note that we do not perform task-specific fine-tuning on any of the evaluation benchmarks. Instead, a single unified model is evaluated directly on all benchmarks for both tasks, demonstrating generalization and robust dual-task alignment. For region captioning, we use DLC-Bench [16] and GAR-Bench-VQA [30] to assess discriminative language capability and descriptive richness. For region localization, we evaluate spatial grounding performance on GCG [22], GRES [18], and the GroundingSuite [9] (covering *Stuff*, *Part*, and *Multi-object* scenarios). GRES is subsequently used to evaluate the model's target-rejection capability.

Comparison Baselines. We compare our CycleGRPO against a comprehensive suite of state-of-the-art models, including: (1) General MLLMs, such as the private GPT-4o [11], o3 [21], and Gemini 2.5 Pro [6], alongside the open-source Qwen2.5-VL [2] and InternVL3 series [59]; (2) Region-level and Segmentation MLLMs, encompassing mask-based models like LISA [12], GLaMM [22], Sa2VA [45], VP-SPHINX [17], EVF-SAM [56], SAMTok, and ARGenSeg [40], as well as specialized referring segmentation models such as SAM4MLLM [5], MLLMSeg [35], and HiMTok [37]; and (3) Task-specific Baselines, including DAM [16] for region understanding and InstructSeg [41] for grounding tasks.

4.1 Main Results

Region Captioning. Table 1 summarizes the performance on DLC-Bench, reported under the "Llama3.1-8B without images" evaluation mode. CycleGRPO

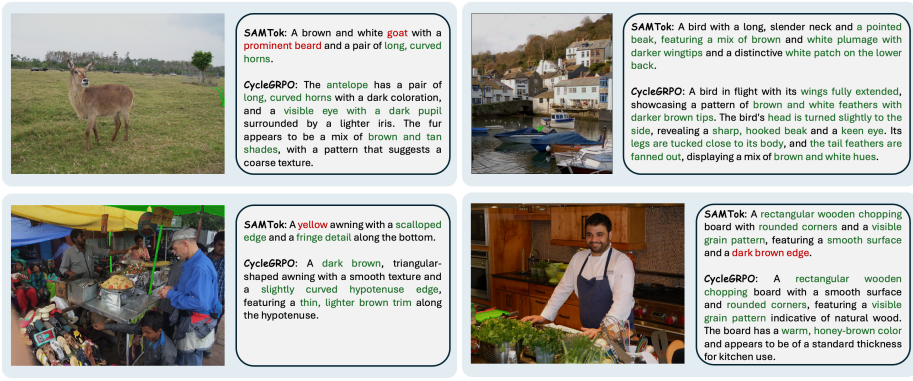

Fig. 3: Qualitative comparison of region captioning on DLC-Bench [16]. We compare the descriptions generated by SAMTok and CycleGRPO for the regions **highlighted in green**. In the text, green indicates accurate attributes and red denotes errors. Compared to SAMTok [57], CycleGRPO produces more **factually accurate** descriptions (e.g., correctly identifying the "antelope" instead of a "goat") and provides significantly **richer details** regarding texture, shape, and color. These cases demonstrate that our cycle-consistency reward effectively encourages the model to distill comprehensive semantic information from visual regions.

boosts the base SAMTok model's average score from 61.9 to 67.7, outperforming closed-source frontiers like GPT-4o (61.5) and Gemini 2.5 Pro (55.8). Notably, it matches DAM (67.3) despite DAM being trained on 1.5M in-domain region-caption pairs that our model never encountered, suggesting that the self-evaluating loop in CycleGRPO can substitute for massive human annotations by bootstrapping from the model's internal spatial-semantic duality. Compared to standard GRPO (63.4), our cycle-consistent approach yields an additional 4.3 gain, indicating that intrinsic spatial grounding provides a more precise supervisory signal than generic RL. These gains are qualitatively reflected in Fig. 3, where CycleGRPO generates more discriminative, factually grounded descriptions and reduces semantic hallucinations by anchoring text to precise spatial reconstructions.

Region VQA. As shown in Table 3, CycleGRPO elevates SAMTok's overall score to 65.1%, surpassing frontier closed-source models such as Gemini-2.5-Pro (64.2%) and o3 (61.3%). The improvement is particularly pronounced in fine-grained *Perception*, where our model delivers consistent gains in categories such as Color (+4.3%), Shape (+1.6%), and Material (+2.8%). Furthermore, in complex *Reasoning*, CycleGRPO achieves a significant 3.3% boost in the Non-Entity split compared to the SAMTok baseline. The slight drops on the Position (-3.2%) and Relation (-1.0%) splits stem from the nature of our reward: mask reconstruction depends more on attribute descriptions (e.g., color, shape) than on spatial relations, steering the model to prioritize the former. Compared to naive GRPO, which shows negligible or even negative impact on this benchmark (63.9% vs.

Table 3: Comparison on GAR-Bench-VQA [31]. This benchmark consists of multiple-choice questions designed to evaluate the model's fine-grained perception and reasoning capabilities. We report the accuracy (%) across various categories, where † denotes models evaluated in thinking mode. Our method demonstrates superior zero-shot performance, particularly in complex perception and non-entity reasoning, even without task-specific tuning or human-annotated VQA data.

Method	Size	Overall	Perception				Reasoning		
			Color	Shape	Texture	Material	Position	Non-Entity	Relation
GPT-4o [11]	-	53.5	34.8	65.3	48.3	52.8	57.8	60.2	61.4
o3† [21]	-	61.3	58.0	70.3	55.2	63.9	54.7	49.2	71.3
Gemini-2.5-Pro† [6]	-	64.2	62.3	68.8	58.6	66.7	64.1	64.9	70.3
Qwen2.5-VL [2]	3B	34.4	29.0	25.0	34.5	30.6	43.8	26.2	44.6
Qwen2.5-VL [2]	7B	41.7	39.1	40.6	44.8	27.8	59.4	36.1	40.6
Qwen2.5-VL [2]	32B	50.9	46.4	53.1	41.4	30.6	71.9	36.1	58.4
Qwen2.5-VL [2]	72B	52.8	46.4	50.0	65.5	33.3	68.8	44.3	57.4
InternVL3 [59]	2B	35.1	30.4	21.9	48.3	38.9	48.4	26.2	38.6
InternVL3 [59]	8B	38.9	36.2	37.5	58.6	41.7	51.6	27.9	33.6
InternVL3 [59]	32B	46.5	39.1	40.6	51.7	55.6	60.9	36.1	47.5
InternVL3 [59]	78B	50.5	44.9	54.7	58.6	61.1	53.1	47.5	45.5
Sa2VA [45]	8B	34.3	39.1	45.3	29.6	30.6	54.7	21.3	21.8
VP-SPHINX [17]	13B	37.5	33.3	25.0	44.8	38.9	60.9	34.3	32.7
DAM [16]	3B	38.2	55.1	39.1	41.4	36.1	31.3	36.1	31.7
GAR [31]	1B	50.6	55.1	46.9	69.0	47.2	21.9	62.3	56.4
GAR [31]	8B	59.9	59.4	54.7	75.9	52.8	48.4	60.7	68.3
SAMTok [57]	4B	64.2	58.0	48.4	48.3	58.3	76.6	54.1	83.2
+ GRPO	4B	63.9	58.0	46.9	44.8	58.3	76.6	54.1	84.2
+ CycleGRPO	4B	65.1	62.3	50.0	48.3	61.1	73.4	57.4	82.2
Δ v.s. SAMTok [57]		↑ 0.9	↑ 4.3	↑ 1.6	– 0.0	↑ 2.8	↓ 3.2	↑ 3.3	↓ 1.0
Δ v.s. GRPO		↑ 1.2	↑ 4.3	↑ 3.1	↑ 3.5	↑ 2.8	↓ 3.2	↑ 3.3	↓ 2.0

64.2%), our cycle-consistency reward provides a more effective training signal by anchoring textual descriptions to physical spatial reconstruction. These results demonstrate that CycleGRPO successfully distills superior visual-reasoning capabilities, achieving better factual grounding without requiring any task-specific tuning or human-annotated VQA data.

Interleaved Text-mask Generation. The GCG benchmark (Table 4) jointly evaluates understanding and grounding by requiring simultaneous generation of descriptive text and precise masks. CycleGRPO consistently outperforms the SAMTok baseline, with a significant leap in captioning quality (+6.5 CIDEr, +1.1 METEOR on Val) alongside improved grounding (+3.0% Recall, +1.2% AP_{50}). Compared to standard GRPO (+3.8 CIDEr), our cycle-consistency objective provides a more effective signal for semantic-spatial synchronization. Notably, despite its 4B scale, our model surpasses larger counterparts like GLaMM (7B) and Sa2VA (8B) in both descriptive richness and localization accuracy.

Text-to-Mask. We evaluate spatial grounding on the GRES and GroundingSuite benchmarks. As shown in Table 5, CycleGRPO achieves a dramatic leap on GRES: the model's ability to reject non-existent targets (Avg. N-acc) soars

Table 4: Results on interleaved text-mask generation task (GCG) [22]. This benchmark provides a holistic evaluation of the model's **joint understanding and grounding** capabilities by requiring the simultaneous generation of descriptive text and precise segmentation masks. We report linguistic metrics (METEOR, CIDEr) alongside spatial metrics (AP_{50}, mIoU, Recall) to assess the semantic-spatial synchronization. CycleGRPO consistently improves both text quality and localization accuracy.

Method	Size	Val					Test				
		METEOR	CIDEr	AP50	mIoU	Recall	METEOR	CIDEr	AP50	mIoU	Recall
LISA [12]	7B	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
GLaMM [22]	7B	16.2	47.2	30.8	66.3	41.8	15.8	43.5	29.2	65.6	40.8
OMG-LLaVA [50]	7B	14.9	41.2	29.9	65.6	–	14.5	38.5	28.6	64.7	–
Sa2VA [45]	8B	16.4	49.5	33.2	67.7	45.1	16.2	49.0	32.2	66.8	44.5
SAMTok [57]	4B	16.1	48.2	34.7	69.4	46.6	16.4	51.4	34.4	68.4	48.3
+ GRPO	4B	16.4	50.9	35.8	69.6	47.5	16.4	52.3	35.3	68.9	47.9
+ CycleGRPO	4B	17.2	54.7	35.9	69.6	49.6	17.1	54.0	35.2	68.6	49.7
Δ *v.s.* SAMTok [57]		↑ 1.1	↑ 6.5	↑ 1.2	↑ 0.2	↑ 3.0	↑ 0.7	↑ 2.6	↑ 0.8	↑ 0.2	↑ 1.4
Δ *v.s.* GRPO		↑ 0.8	↑ 3.8	↑ 0.1	- 0.0	↑ 2.1	↑ 0.7	↑ 1.7	↓ 0.1	↓ 0.3	↑ 1.8

Table 5: Results on text-to-mask task (GRES) [18]. This benchmark evaluates Referring Expression Segmentation (RES) and target rejection capability. We report gIoU, cIoU, and N-acc (%).

Method	Size	Val			Test A			Test B			Avg.		
		gIoU	cIoU	N-acc	gIoU	cIoU	N-acc	gIoU	cIoU	N-acc	gIoU	cIoU	N-acc
LISA [12]	7B	61.6	61.8	54.7	66.3	68.5	50.0	58.8	60.6	51.9	62.2	63.6	52.2
SAM4MLLM [5]	8B	71.9	67.8	66.1	74.2	72.2	63.9	65.3	63.4	60.0	70.5	67.8	63.3
MLLMSeg [35]	8B	75.1	71.6	73.2	77.0	76.9	72.4	69.7	68.5	65.5	73.9	72.3	70.4
HiMTok [37]	8B	72.1	70.4	–	73.5	74.9	–	71.7	72.0	–	72.4	72.4	–
ARGenSeg [40]	8B	74.7	72.2	–	73.7	73.6	–	72.4	70.4	–	73.6	72.1	–
SAMTok [57]	4B	71.3	69.2	61.4	75.3	75.4	59.0	66.9	66.0	55.6	71.2	70.2	58.7
+ GRPO	4B	75.7	68.3	98.8	70.2	69.8	98.1	66.4	64.0	96.8	70.8	67.4	97.9
+ CycleGRPO	4B	81.8	74.6	94.2	79.9	77.8	93.1	73.0	70.0	89.0	78.2	74.1	92.1
Δ *v.s.* SAMTok [57]		↑ 10.5	↑ 5.4	↑ 32.8	↑ 4.6	↑ 2.4	↑ 34.1	↑ 6.1	↑ 4.0	↑ 33.4	↑ 7.0	↑ 3.9	↑ 33.4
Δ *v.s.* GRPO		↑ 6.1	↑ 6.3	↓ 4.6	↑ 9.7	↑ 8.0	↓ 5.0	↑ 6.6	↑ 6.0	↓ 7.8	↑ 7.4	↑ 6.7	↓ 5.8

from 58.7% to 92.1%, indicating that our cyclic objective curbs hallucinations by penalizing captions that lead to failed spatial reconstructions. It also delivers a 7.0% gain in average gIoU over SAMTok, outperforming dedicated segmentation models such as LISA (62.2%) and ARGenSeg (73.6%). On GroundingSuite (Table 2), CycleGRPO lifts overall accuracy from 57.5% to 67.6%, with substantial gains on Multi-object (+14.3%) and Single-object (+8.7%) tasks. These improvements are visually evident in Fig. 4 and Fig. 5.

4.2 Ablation and Analysis

Comparison with GRPO baselines. To verify the necessity of the cycle-consistency reward, we compare CycleGRPO with various GRPO baselines. We first prompt the MLLM to generate captions for given regions. These generated

Fig. 4: Qualitative results on GRES [18]. We compare CycleGRPO with SAMTok [57] and the ground truth. These examples require advanced spatial reasoning to distinguish targets in complex scenes. In the first row, CycleGRPO accurately localizes multiple disjoint entities mentioned in the query, whereas SAMTok incorrectly includes the man's coat. In the second row, our model correctly identifies the target despite occlusions, while the baseline fails to provide a valid segmentation.

Table 6: Ablation on per-task GRPO baselines on SAMTok pretrained on Qwen3-VL-4B. Each single-direction variant (Captioning-only or Localization-only GRPO) improves only its own task, while CycleGRPO improves both and surpasses each baseline on its home metric.

Method	DLC-Bench			GroundingSuite				
	Pos.	Neg.	Avg.	Single	Stuff	Multi	Part	All
SAMTok-4B	43.5	80.4	61.9	80.9	12.4	62.0	52.9	57.5
+Cap-Only GRPO	46.7	**85.6**	66.2	80.7	11.9	62.2	51.7	56.9
+Loc-Only GRPO	43.2	80.4	61.8	90.7	18.4	67.7	60.3	65.0
+CycleGRPO	**50.6**	85.4	**68.0**	**91.5**	**24.2**	**74.5**	**63.4**	**68.5**

captions are then used as text prompts for a subsequent referring segmentation task, where the resulting IoU serves as the reward signal for policy optimization.

The fundamental difference lies in the optimization target: while the standard GRPO baseline utilizes the spatial reconstruction accuracy (IoU) to tune the model, it lacks the direct, end-to-end incentive to optimize the semantic discriminability of the generated captions themselves. As shown in Table 1, although standard GRPO achieves moderate gains on DLC-Bench (+1.5 Avg. score), it is significantly outperformed by CycleGRPO (+5.8 Avg. score). Furthermore, on the GRES benchmark (Table 5), standard GRPO even suffers from a slight regression in segmentation precision (71.2 → 70.8 IoU). These results demonstrate that without the "region → text → region" loop, the model tends to fall into a sub-optimal policy that sacrifices descriptive richness for easier spatial matching.

We further compare with separate per-task (captioning-only or grounding-only) GRPO baselines, each of which optimizes a single direction and requires

Fig. 5: Qualitative results on GroundingSuite [9]. We compare our CycleGRPO with SAMTok [57] and the ground truth. The first row demonstrates **part-level segmentation** (e.g., the duck's head), which requires fine-grained understanding of object components. The second row illustrates **stuff-class segmentation** (e.g., the snow), highlighting the model's capability for context-aware localization in amorphous regions. CycleGRPO consistently produces masks that are more precisely aligned with the textual queries.

GT captions either as supervision or as input. Our training data contains only (image, mask) pairs without caption GT, precluding direct per-task baselines on it. We thus sampled (image, mask, GT caption) triples from the DAM training set [16] and trained both per-task baselines and CycleGRPO under matched compute, following Fig. 1. Captioning-only GRPO (image + mask $\rightarrow$ caption) uses an LLM-as-judge with Qwen2.5-72B on (predicted caption, GT caption) pairs; Localization-only GRPO (image + caption $\rightarrow$ mask) uses IoU as the reward. As shown in Tab. 6, each per-task baseline improves its own task but slightly drops the other. CycleGRPO, without using the GT caption, improves both and surpasses each baseline on its home metric, demonstrating mutual benefit from the cycle structure.

Generalizability Across Grounding Formats. To investigate whether the efficacy of CycleGRPO is restricted to the SAMTok-based mask discretization, we conduct additional experiments using bounding boxes (bbox) as the spatial grounding format. In this setting, we utilize Qwen2.5VL-3B [2] and Qwen3VL-4B&8B [1] as the base models, where region prompts for captioning and outputs for localization are represented in the standard bounding box format (e.g., `[y_min, x_min, y_max, x_max]`) instead of discrete mask tokens.

As shown in Table 8, CycleGRPO consistently improves performance across vanilla MLLM backbones of varying scales, validating its generalizability beyond specific segmentation tokenizers like SAMTok. When applied to the bbox-based setting, CycleGRPO yields steady gains on DLC-Bench across all three backbones, lifting the average score by 2.9, 3.3, and 4.0 points on Qwen2.5-VL-3B, Qwen3-VL-4B, and Qwen3-VL-8B, respectively. The benefits extend to spatial grounding as well: on GroundingSuite (Acc@0.5), CycleGRPO im-

Table 7: Ablation on the number of rollouts G and K. We evaluate the impact of varying the captioning group size G and localization rollout count K on DLC-Bench and GroundingSuite. The results show that increasing the sampling density generally improves performance by providing more stable feedback. We select $G = K = 6$ as our default configuration to maintain an optimal trade-off between training efficacy and computational cost.

Rollout	DLC-Bench			GroundingSuite				
	Pos.	Neg.	Avg.	Stuff	Part	Multi	Single	All
G=K=0 (SAMTok)	43.5	80.4	61.9	80.9	12.4	62.0	52.9	57.5
G=K=2	45.8	82.0	63.9	87.6	12.2	74.1	54.9	62.5
G=K=6	51.2	84.2	67.7	90.7	20.9	76.3	61.6	67.6
G=K=10	51.1	84.8	67.9	91.1	25.4	76.7	63.7	69.2

Table 8: Ablation on spatial grounding formats on vanilla MLLM backbones. We evaluate the generalizability of CycleGRPO by replacing the discrete mask tokens with continuous bounding box coordinates.

Method	DLC-Bench			GroundingSuite (Acc@0.5)				
	Pos.	Neg.	Avg.	Single	Stuff	Multi	Part	All
Qwen2.5-VL-3B (bbox)	19.4	44.8	32.1	93.7	32.8	63.2	66.3	69.0
+ CycleGRPO	**23.3**	**46.6**	**35.0**	**94.1**	**34.1**	**64.1**	**67.3**	**69.9**
Qwen3-VL-4B (bbox)	39.6	62.4	51.0	**95.2**	24.2	**76.7**	67.0	71.4
+ CycleGRPO	**43.2**	**65.4**	**54.3**	94.9	**33.9**	76.3	**71.1**	**74.1**
Qwen3-VL-8B (bbox)	41.0	**65.4**	53.2	94.9	31.2	71.8	67.4	71.4
+ CycleGRPO	**49.4**	65.0	**57.2**	**95.9**	**37.6**	**74.4**	**70.4**	**74.1**

proves the overall score from 71.4 to 74.1 on both Qwen3-VL-4B and Qwen3-VL-8B. These results underscore that CycleGRPO is a generalizable framework that reinforces the spatial-semantic closed-loop across different model scales and grounding modalities, effectively scaling the model's fine-grained understanding regardless of the underlying spatial representation.

Impact of Rollout Count G and K. We investigate the sensitivity of CycleGRPO to the hyperparameters G and K. As shown in Table 7, increasing the sampling density generally leads to more stable feedback and better performance. Specifically, as G and K increase from 2 to 6, the average score on DLC-Bench rises from 61.9% to 67.7%, while the overall accuracy on GroundingSuite jumps from 57.5% to 67.6%. Notably, the performance on the challenging Part split nearly doubles, reaching 20.9%. Although further scaling to $G = K = 10$ still yields slight additional gains (e.g., 69.2% on GroundingSuite), we select $G = K = 6$ as our default configuration to maintain an optimal trade-off between computational cost and training efficacy. Additional ablation studies are provided in the supplementary materials.

5 Related Work

Region Understanding for MLLMs. Conventional MLLMs have demonstrated remarkable capabilities in image-level holistic understanding [2, 19, 21, 29, 32–34, 38, 42]. To achieve a much more fine-grained visual comprehension, region-aware MLLMs have been proposed [7, 16, 31, 46–49, 57]. However, most of these methods rely heavily on *high-quality SFT data*, *e.g.*, detailed region captions, which incur high annotation costs and are poorly scalable in real-world scenarios. To address this, we present the *actor as its own critic*, a scalable cycle reinforcement learning pipeline that requires only region inputs. By leveraging cycle-consistent RL rewards, *i.e.*, reconstructive IoU, the model is incentivized to generate discriminative descriptions of similar regions that capture unique details, thereby facilitating precise region understanding and localization.

Visual Region Localization for MLLMs. Precise region localization is a core capability of modern MLLMs [1, 8, 10, 28, 38]. Existing post-training approaches [12, 13, 25, 36, 37, 39, 40, 45, 50–52, 57] typically enhance this ability by using instruction-based visual grounding dialogues that rely on large-scale, high-quality expression–location annotations. Such dependence creates a data bottleneck that limits further performance gains. We address this issue with a CycleGRPO framework that requires only region locations as supervision.

RL for Region Understanding and Localization. In RL post-training for region understanding and localization, prior works [29, 43, 57] typically optimize grounding and captioning separately. For visual grounding [3, 4, 20, 23, 44, 58, 60], the rewards are calculated from distances or IoU between predicted locations and ground truth, which requires high-quality expression–location pairs. For region captioning [16, 26, 31], reward design is more complex: methods such as CapRL [43] evaluate captions indirectly via question answering, relying on curated QA pairs and additional LLMs for reward computation, thereby increasing data and system costs. In contrast, our framework unifies two tasks under a shared objective, simplifies the training pipeline, and reduces reliance on auxiliary data and models.

6 Conclusion

This paper presented **CycleGRPO**, a reinforcement learning framework that leverages spatial-semantic duality to bootstrap multimodal capabilities without human-annotated region-caption pairs. By framing region understanding and localization as a self-evaluating closed loop, our method derives training signals from the cycle consistency between generated text and reconstructed spatial regions. Extensive experiments show that CycleGRPO significantly enhances both understanding and grounding across multiple benchmarks, outperforming much larger closed-source and open-source frontiers and generalizing to different spatial representations such as continuous bounding boxes. We argue that the "Actor-as-Own-Critic" paradigm offers a scalable path for the continued evolution of multimodal models without expensive human data.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. 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)
3. Bai, S., Li, M., Liu, Y., Tang, J., Zhang, H., Sun, L., Chu, X., Tang, Y.: Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. arXiv preprint arXiv:2505.14231 (2025)
4. Cao, M., Zhao, H., Zhang, C., Chang, X., Reid, I., Liang, X.: Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. arXiv preprint arXiv:2505.20272 (2025)
5. 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: ECCV (2024)
6. DeepMind, G.: Gemini-2.5-pro. https://deepmind.google/models/gemini/pro/ (2025)
7. Guo, Q., De Mello, S., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. In: CVPR. pp. 13796–13806 (2024)
8. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024)
9. Hu, R., Zhu, L., Zhang, Y., Cheng, T., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., Wang, X.: Groundingsuite: Measuring complex multi-granular pixel grounding. arXiv preprint arXiv:2503.10596 (2025)
10. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., et al.: Step3-vl-10b technical report. arXiv preprint arXiv:2601.09668 (2026)
11. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
12. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR (2024)
13. Lan, M., Chen, C., Zhou, Y., Xu, J., Ke, Y., Wang, X., Feng, L., Zhang, W.: Text4seg: Reimagining image segmentation as text generation. ICLR (2025)
14. Li, R., Li, L., Ren, S., Tian, H., Gu, S., Li, S., Yue, Z., Wang, Y., Ma, W., Yang, Z., et al.: Groundingme: Exposing the visual grounding gap in mllms through multi-dimensional evaluation. arXiv preprint arXiv:2512.17495 (2025)
15. Li, X., Zhang, T., Li, Y., Yuan, H., Chen, S., Zhou, Y., Meng, J., Sun, Y., Xu, S., Qi, L., et al.: Denseworld-1m: Towards detailed dense grounded caption in the real world. arXiv preprint arXiv:2506.24102 (2025)
16. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video captioning. arXiv preprint arXiv:2504.16072 (2025)
17. Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. arXiv preprint arXiv:2403.20271 (2024)

18. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR (2023)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
20. 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)
21. OpenAI: Openai-o3. https://openai.com/index/introducing-o3-and-o4-mini/ (2025)
22. 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: CVPR (2024)
23. Sarch, G., Saha, S., Khandelwal, N., Jain, A., Tarr, M.J., Kumar, A., Fragkiadaki, K.: Grounded reinforcement learning for visual reasoning. arXiv preprint arXiv:2505.23678 (2025)
24. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
25. Su, Y., Zhang, H., Li, S., Liu, N., Liao, J., Pan, J., Liu, Y., Xing, X., Sun, C., Li, C., et al.: Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. arXiv preprint arXiv:2510.01954 (2025)
26. Sun, Y., Wang, Y., Li, J., Tian, Y., Zhang, T., Mai, J., Wang, Y., Wang, H., Bai, J., Yang, L., Tong, Y.: Perceptiondlm: Parallel region perception with multimodal diffusion language models. arXiv preprint arXiv:2606.19534 (2026), https://arxiv.org/abs/2606.19534
27. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
28. Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., Tang, J.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2024)
29. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., et al.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. ICLR (2026)
30. Wang, H., Wang, Y., Zhang, T., Zhou, Y., Li, Y., Wang, J., Tian, Y., Meng, J., Huang, Z., Mai, G., Wang, A., Tong, Y., Wang, Z., Li, X., Zhang, Z.: Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. ICLR (2026)
31. Wang, H., Wang, Y., Zhang, T., Zhou, Y., Li, Y., Wang, J., Tian, Y., Meng, J., Huang, Z., Mai, G., et al.: Grasp any region: Towards precise, contextual pixel understanding for multimodal llms. arXiv preprint arXiv:2510.18876 (2025)

32. Wang, H., Zhao, Y., Wang, T., Fan, H., Zhang, X., Zhang, Z.: Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9275–9286 (2025)
33. Wang, H., Zheng, A., Zhao, Y., Wang, T., Ge, Z., Zhang, X., Zhang, Z.: Reconstructive visual instruction tuning. In: ICLR (2025)
34. Wang, J., Kang, Z., Wang, H., Jiang, H., Li, J., Wu, B., Wang, Y., Ran, J., Liang, X., Feng, C., et al.: Vgr: Visual grounded reasoning. arXiv preprint arXiv:2506.11991 (2025)
35. Wang, J., Wu, Z., Huang, D., Zheng, Y., Wang, H.: Unlocking the potential of mllms in referring expression segmentation via a light-weight mask decoder. arXiv preprint arXiv:2508.04107 (2025)
36. Wang, L., Lin, H., Chen, S., Wang, T., Cheng, C., Zhong, Y., Zheng, D., Zhao, W.: Alto: Adaptive-length tokenizer for autoregressive mask generation. arXiv preprint arXiv:2505.16495 (2025)
37. Wang, T., Cheng, C., Wang, L., Chen, S., Zhao, W.: Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In: ICCV (2025)
38. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
39. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. NeurIPS (2023)
40. Wang, X., Ru, L., Huang, Z., Ji, K., Zheng, D., Chen, J., Zhou, J.: Argenseg: Image segmentation with autoregressive image generation model. In: NeurIPS (2025)
41. Wei, C., Zhong, Y., Tan, H., Zeng, Y., Liu, Y., Wang, H., Yang, Y.: Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In: ICCV. pp. 20193–20203 (2025)
42. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
43. Xing, L., Dong, X., Zang, Y., Cao, Y., Liang, J., Huang, Q., Wang, J., Wu, F., Lin, D.: Caprl: Stimulating dense image caption capabilities via reinforcement learning. arXiv preprint arXiv:2509.22647 (2025)
44. You, Z., Wu, Z.: Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. arXiv preprint arXiv:2506.22624 (2025)
45. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
46. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: CVPR (2024)
47. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: Videorefer suite: Advancing spatial-temporal object understanding with video llm. In: CVPR (2025)
48. Yuan, Y., Zhang, W., Li, X., Wang, S., Li, K., Li, W., Xiao, J., Zhang, L., Ooi, B.C.: Pixelrefer: A unified framework for spatio-temporal object referring with arbitrary granularity. arXiv preprint arXiv:2510.23603 (2025)
49. Zhang, H., You, H., Dufter, P., Zhang, B., Chen, C., Chen, H.Y., Fu, T.J., Wang, W.Y., Chang, S.F., Gan, Z., et al.: Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:2404.07973 (2024)

50. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C.C., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. NeurIPS (2024)
51. Zhang, T., Li, X., Huang, Z., Li, Y., Lei, W., Deng, X., Chen, S., Ji, S., Feng, J.: Pixel-sail: Single transformer for pixel-grounded understanding. arXiv preprint arXiv:2504.10465 (2025)
52. Zhang, T., Wei, S., Chen, S., Yu, W., Luo, M., Ji, S.: Vectorllm: Human-like extraction of structured building contours vis multimodal llms. arXiv preprint arXiv:2507.04664 (2025)
53. Zhang, X., Chen, Y.C.: Adaptive domain generalization via online disagreement minimization. IEEE Transactions on Image Processing **32**, 4247–4258 (2023)
54. Zhang, X., Tan, R.T.: Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14527–14537 (2025)
55. Zhang, X., Xie, J., Yuan, Y., Mi, M.B., Tan, R.T.: Heap: unsupervised object discovery and localization with contrastive grouping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7323–7331 (2024)
56. Zhang, Y., Cheng, T., Zhu, L., Hu, R., Liu, L., Liu, H., Ran, L., Chen, X., Liu, W., Wang, X.: Evf-sam: Early vision-language fusion for text-prompted segment anything model. arXiv preprint arXiv:2406.20076 (2024)
57. Zhou, Y., Zhang, T., Gong, D., Wu, Y., Tian, Y., Wang, H., Yuan, H., Wang, J., Qi, L., Fei, H., et al.: Samtok: Representing any mask with two words. arXiv preprint arXiv:2601.16093 (2026)
58. Zhou, Y., Dai, S., Wang, S., Zhou, K., Jia, Q., Xu, J.: Gui-g1: Understanding r1-zero-like training for visual grounding in gui agents. arXiv preprint arXiv:2505.15810 (2025)
59. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
60. 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)