

---

# High-Resolution Visual Reasoning via Multi-Turn Grounding-Based Reinforcement Learning

---

Xinyu Huang<sup>1,2</sup> Yuhao Dong<sup>2</sup> Wei Li<sup>3</sup> Jinming Wu<sup>3</sup> Zihao Deng<sup>3</sup>

Weiwei Tian<sup>1</sup> Bo Li<sup>2</sup> Rui Feng<sup>1</sup> Zejun Ma<sup>3</sup> Ziwei Liu<sup>2</sup>

<sup>1</sup>Fudan University <sup>2</sup>S-Lab, Nanyang Technological University <sup>3</sup>TikTok

## Abstract

State-of-the-art large multi-modal models (LMMs) face challenges when processing high-resolution images, as these inputs are converted into enormous visual tokens, many of which are irrelevant to the downstream task. In this paper, we propose **Multi-turn Grounding-based Policy Optimization (MGPO)**, an end-to-end reinforcement learning (RL) framework that enables LMMs to iteratively focus on key visual regions by automatically cropping sub-images, based on model-predicted grounding coordinates within a multi-turn conversation framework. Compared to supervised fine-tuning (SFT), which requires costly additional grounding annotations, *our approach highlights that LMMs can emerge robust grounding abilities during the RL training process, leveraging only a binary reward function derived from the correctness of the final answer.* Additionally, we observe that LMMs struggle to autonomously trigger visual grounding during the rollout process. To address this cold start problem, we design a multi-turn conversational template and restrict policy loss computation to model outputs generated across multiple dialogue rounds, thereby promoting stable optimization. Extensive experiments demonstrate that, when trained on standard visual-question-short answering data without grounding annotations, MGPO effectively elicits stronger grounding capabilities compared to GRPO, leading to 5.4% improvement on in-distribution MME-Realworld and 5.2% improvement on the challenging out-of-distribution (OOD) V\* Bench. Notably, MGPO post-training on Qwen2.5-VL-7B with 21K samples surpasses OpenAI’s o1 and GPT-4o models on the OOD V\* Bench.

## 1 Introduction

State-of-the-art large multimodal model (LMM) architectures, such as Qwen2.5-VL [1], typically build on a powerful large language model (LLM) backbone (e.g., Qwen2.5 [43]) integrated with an external Native Resolution Vision Transformer (NaViT) module [6, 39]. The NaViT module enables LMMs to process images at their native resolutions, allowing the model to clearly “see” the details of images and achieve high perceptual fidelity [17, 20].

Nevertheless, such approach also presents challenges in high-resolution real-world scenarios. (1) The number of visual tokens increases quadratically with image resolution, resulting in a large proportion of tokens that may be irrelevant to the downstream task—an issue analogous to the “needle-in-a-haystack” problem in LLMs [21]. (2) Due to the inherent context length limitations of LMMs, a maximum pixel constraint is imposed on input images in practical applications, necessitating to resize images that exceed this threshold.

Drawing inspiration from the human visual system, while its theoretical resolving power is estimated to be equivalent to approximately 576 megapixels [4, 5], high-acuity vision is restricted to the foveal region of the macula, which covers only about 125,000 pixels within a central 1° visual angle [4, 36]. As a result, when processing high-resolution real-world scenarios, the human visual system employs

task-driven visual search strategies [41] to ground and scrutinize critical regions of interest. Motivated by this biological mechanism, we attempt to equip LLMs with similar visual search capabilities by leveraging visual grounding to focus on key image regions.

However, empowering LMMs with such grounding-based visual reasoning capabilities is non-trivial, primarily due to the scarcity and high cost of obtaining grounding annotations for standard visual-question-answering (VQA) datasets, which are required for constructing multi-turn grounding-based conversation data for supervised fine-tuning (SFT). In this paper, *we highlight that accurate grounding behavior can emerge within a reinforcement learning (RL) paradigm, even when training supervision is provided solely through a binary reward function derived from the correctness of the final answer.*

To this end, we introduce **Multi-turn Grounding-based Policy Optimization (MGPO)**, a reinforcement learning (RL) algorithm that enables LMMs to iteratively focus on key image regions by automatically cropping sub-images, based on model-predicted grounding coordinates within a multi-turn conversation framework. Given a high-resolution image and a question, the model first predicts the coordinates of key regions relevant to the query. An image cropping function is then triggered to extract and return the corresponding sub-image. In subsequent turns, the model can integrate previous in-context conversations (including both the original image and cropped sub-image) to solve the question.

Empirically, we observe that LMMs struggle to autonomously trigger visual grounding during the rollout process. To mitigate the cold start problem without constructing additional annotated data, we design a fixed multi-turn conversation template: the first turn prompts the model to output relevant coordinates, while the second turn provides the sub-image and prompts the model to answer the question. To ensure stable optimization, policy loss is computed only on model outputs generated across multiple conversation rounds.

In summary, MGPO mainly offers the following advantages:

- **Top-down and Interpretable Visual Reasoning.** MGPO equips LMMs with a top-down, question-driven visual search mechanism for high-resolution scenarios and provides interpretable outputs that indicate which image regions are attended to throughout the reasoning process.
- **Overcomes Maximum Pixel Constraints.** MGPO can overcomes the maximum pixel limitation of LMMs. As shown in the first example of Figure 1, even when resizing a high-resolution image within pixel limits results in a blurred input, the model can still identify relevant coordinates and crop clear sub-images from the original input for further analysis.
- **Without Additional Grounding Annotations.** MGPO can be post-trained directly on standard VQA datasets without the need for extra grounding annotations, and experimental results demonstrate substantial improvements in intermediate grounding performance compared to GRPO [33].

Ultimately, we utilize MGPO to post-train Qwen2.5-VL-7B [1] using visual-question-short answering data, yet achieves strong intermediate grounding performance without requiring grounding annotations (examples shown in Figure 1). Compared to GRPO, MGPO yields a 5.4% improvement on the in-distribution MME-Realworld [49] benchmark and a 5.2% gain on the challenging out-of-distribution (OOD) V\* Bench [41]. Notably, leveraging with only 21K post-training samples, our model surpasses OpenAI’s o1 and GPT-4o models on the OOD V\* Bench.

## 2 Preliminaries

### 2.1 Advanced Baseline LMMs

As our baseline, we adopt Qwen2.5-VL [1], an advanced LMM that combines the Qwen2.5 [43] large language model with a Native Resolution Vision Transformer (NaViT) [6, 39]. The NaViT module enables the model to process images at their native resolutions by dividing the input image  $I$  of height  $h$  and width  $w$  into non-overlapping patches of size  $p \times p$ . Groups of  $m \times m$  patches are then aggregated and transformed by a multi-layer perceptron (MLP) to produce a single visual token:

$$X_i = \text{MLP}(\text{Merge}(\text{Patch}(I, p), m)), \quad (1)$$

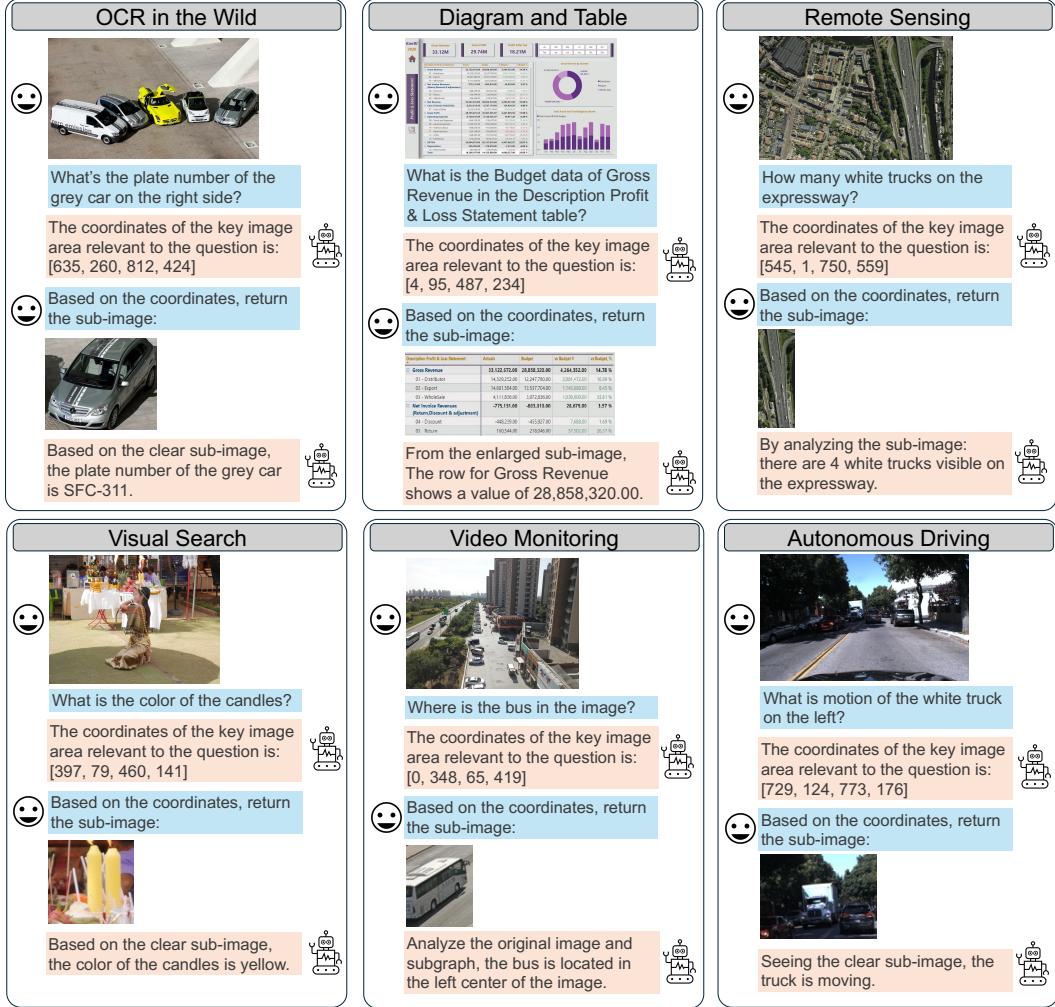


Figure 1: Examples of models trained with multi-turn grounding-based RL on high-resolution real-world tasks. The model first identifies key regions, which are then automatically cropped and returned as sub-images. Notably, despite only a binary reward function derived from the correctness of the final answer, the model gradually emerge robust grounding capability throughout the RL process. The conversation in the figure only shows key parts, the full conversation is provided in Appendix 9.

Visual tokens  $X_i$  are concatenated with textual tokens  $X_t$  and jointly processed by the LLM for multimodal tasks. In addition, Qwen2.5-VL supports visual grounding by outputting bounding box coordinates  $[x_1, y_1, x_2, y_2]$  within the image dimensions  $h$  and  $w$ . However, the model typically requires explicit prompting to produce grounding coordinates, which does not spontaneously invoke grounding as part of its reasoning process.

## 2.2 Single-Turn Multi-modal RL

Figure 2 illustrates a comparison of different post-training paradigms for LMMs. In supervised fine-tuning (SFT), the LMM is trained to directly imitate ground-truth as the response answer. In contrast, reinforcement learning (RL) enables the model to explore effective reasoning trajectories leading to correct answers, thereby potentially beyond the limitations of pure imitation learning.

In this study, we adopt the recently widely acknowledged Group Relative Policy Optimization (GRPO) [33] algorithm as the RL baseline. Formally, given visual tokens  $X_i$  and textual context  $X_t$ , the LMM generates a group of  $G$  candidate outputs denoted as  $\{(X_a^g)\}_{g=1}^G$ , where  $X_a^g$  includes the

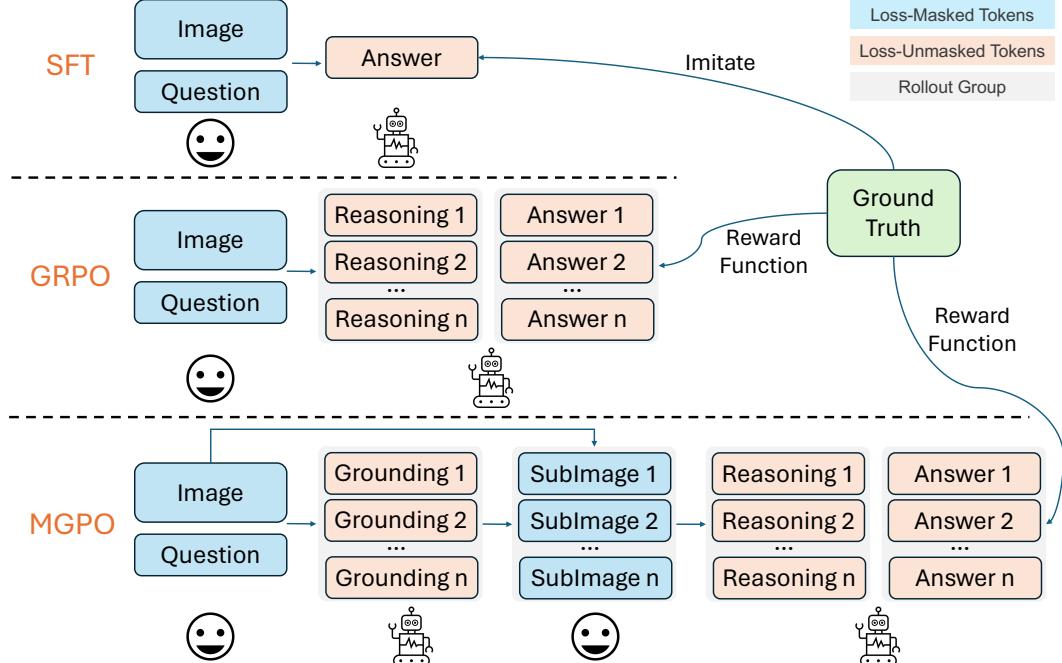


Figure 2: Comparison of different post-training paradigms for LMMs. Our MGPO automatically crops and returns sub-image to the model based on its predicted grounding coordinates, enabling the model to iteratively focus on key regions and effectively solve high-resolution visual tasks.

$g$ -th reasoning trajectory and corresponding answer. The optimization objective is to maximize the likelihood of producing correct answers while minimizing the probability of incorrect ones.

To this end, a rule-based reward function  $\mathcal{R}$  is applied to each candidate answer  $X_a^g$ , yielding a binary reward  $r^g = \mathcal{R}(X_a^g) \in \{0, 1\}$ , where  $r^g = 1$  indicates a correct answer and  $r^g = 0$  indicates an incorrect answer. The policy gradient for GRPO is then formulated as:

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{X_a^g \sim p_{\theta}} [(r^g - b) \cdot \nabla_{\theta} \log p_{\theta}(X_a^g | X_i, X_t)], \quad (2)$$

where  $b = \frac{1}{G} \sum_{g=1}^G r^g$  defined as the average reward within the group, ensuring correct answers receive positive gradient and incorrect receive negative ones. Particularly, only the output tokens  $X_a^g$  are directly involved in the loss calculation, while the input tokens  $X_i$  and  $X_t$  serve solely as conditioning context and are masked out from the loss calculation

### 3 Multi-turn Grounding-Based RL

#### 3.1 Formulation

To address the challenges of high-resolution visual reasoning, we propose Multi-turn Grounding-based Policy Optimization (MGPO). Figure 2 presents the two-turn illustration, and Algorithm 1 provides the general formulation. In this paradigm, the model operates over  $K$  sequential interaction, dynamically grounding and reasoning by conditioning on the full history of visual and textual context at each step.

At each turn  $k$ , the model generates an output  $X_a^{(k),g}$  based on the complete interaction history  $\mathcal{H}^{(k)}$ . If the model's output  $X_a^{(k),g}$  contains grounding coordinates, an image cropping function is triggered to crop the relevant sub-image from the current visual input  $X_i^{(k)}$ , and this sub-image  $X_i^{(k+1)}$  is provided as the new visual input for the next turn. Both the new sub-image and the model's output are appended to the history. The process continues until the model outputs a final answer, at which point the rollout terminates.

---

**Algorithm 1:** Multi-turn Grounding-based Policy Optimization (MGPO)

---

**Input:** Policy model  $\pi_\theta$ ; group size  $G$

**for** each rollout  $g = 1, \dots, G$  **do**

- Initialize  $X_i^{(1)}, X_t^{(1)}$  from image and question;
- $k \leftarrow 1$ ;
- $\mathcal{H}^{(1)} = \{(X_i^{(1)}, X_t^{(1)})\}$ ;
- while** true **do**

  - Sample  $X_a^{(k),g} \sim \pi_\theta(\cdot | \mathcal{H}^{(k)})$ ;
  - if**  $X_a^{(k),g}$  contains final answer **then**

    - break**

  - if**  $X_a^{(k),g}$  contains grounding coordinates **then**

    - $X_i^{(k+1)} \leftarrow \text{Crop}(X_i^{(k)}, \text{Coord}(X_a^{(k),g}))$ ;
    - $\mathcal{H}^{(k+1)} = \mathcal{H}^{(k)} \cup \{X_a^{(k),g}, (X_i^{(k+1)})\}$ ;

  - else**

    - $\mathcal{H}^{(k+1)} = \mathcal{H}^{(k)} \cup \{X_a^{(k),g}\}$ ;

  - $k \leftarrow k + 1$

- Collect  $\{X_a^{(j),g}\}_{j=1}^k$ ;

Compute reward  $r^g = \mathcal{R}(X_a^{(k),g})$  for each rollout  $g$ ;

Compute group average reward  $b = \frac{1}{G} \sum_{g=1}^G r^g$ ;

Update  $\pi_\theta$  using policy gradients calculated by Eq. 3

---

For each question, a group of  $G$  rollouts is sampled. The reward for each rollout is computed based on the final answer, and a group baseline is used to reduce variance. The policy is optimized by maximizing the following objective:

$$\nabla_\theta \mathcal{J}_{\text{MGPO}}(\theta) = \mathbb{E}_{\{X_a^{(j),g}\} \sim p_\theta} \left[ (r^g - b) \cdot \sum_{j=1}^{k_g} \nabla_\theta \log p_\theta \left( X_a^{(j),g} | \mathcal{H}^{(j)} \right) \right], \quad (3)$$

where  $k_g$  denotes the number of steps in the  $g$ -th rollout,  $r^g$  is the reward assigned to the final answer,  $b$  is the average reward within the group, and  $\mathcal{H}^{(j)}$  represents the complete interaction history up to step  $j$ . This formulation encourages the model to optimize all intermediate grounding and reasoning steps that ultimately contribute to a correct final answer.

### 3.2 Implementation Details

**Multi-turn Template without Cold Start.** In practice, we observe that LLMs struggle to autonomously generate grounding coordinates during the rollout process, which hinder effective multi-turn RL. To address this, we design a fixed two-turn dialogue template, as shown in Figure 3, to explicitly activate the model’s grounding and reasoning abilities.

In the first turn, the model is prompted to output only the coordinates of the key image region relevant to the question. In the second turn, the image cropping function first checks the validity of the provided coordinates: if the coordinates are valid, the cropped sub-image is returned to the model; otherwise, the original image is returned. This template-based approach eliminate the cold start, which need to construct multi-turn grounding data and perform SFT before RL.

**Grounding Key Visual Areas.** Within the two-turn MGPO framework, sub-images are extracted directly from the original high-resolution image. This approach is particularly crucial when the original image resolution exceeds the maximum pixel limit of the LMM, as it enables the model to access higher-fidelity sub-image for processing.

Since the grounding coordinates predicted by Qwen2.5-VL are dependent on the resolution of the input image, it is necessary to normalize the predicted coordinates by the input image dimensions

```

User:
<|image_token|>
Question: {question} Out the coordinates of the key image area relevant to the problem in JSON format.

Assistant:
Json
[
    {"bbox_2d": [xxx, xxx, xxx, xxx], "label": xxx}
]

User (Valid Coordinates is Provided):
Based on the provided coordinates, return the sub-image:
<|image_token|>
Question: {question} {choices} Carefully analyze both the original image and the enlarged sub-image to solve the question step by step. If the sub-image does not provide sufficient information, refer to the original image to find the solution. Present your reasoning clearly, and provide the final answer (A, B, C, D, or E) enclosed within \boxed{}.

User (No / Invalid Coordinates is Provided):
Not providing correct coordinates or invalid coordinate values, return the original image:
<|image_token|>
Question: {question} {choices} Carefully analyze the original image to solve the question step by step. Present your reasoning clearly, and provide the final answer (A, B, C, D, or E) enclosed within \boxed{}.

Assistant:
To solve this problem,
...
Based on the analysis, the answer is \boxed{xxx}.

```

Figure 3: Fixed multi-turn grounding template, which eliminate cold start SFT process.

and subsequently map them back to the coordinate space of the original image:

$$[\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2] = \frac{[x_1, y_1, x_2, y_2]}{S_{\text{input}}} \cdot S_{\text{ori}}, \quad (4)$$

where  $S_{\text{input}}$  and  $S_{\text{ori}}$  represent the width and height of the input and original images, respectively. A illustration of this process is provided in the Appendix 8.

## 4 Experiments

### 4.1 Datasets & Metrics

To evaluate the effectiveness of the our approach, experiments are conducted on two established datasets: MME-Realworld [49] and V\* Bench [41]. Both datasets are specifically designed to evaluate the capabilities of large multi-modal (LMMs) in analyzing high-resolution images and capturing fine-grained visual information.

**MME-Realworld.** [49] The MME-Realworld dataset comprises a diverse array of tasks, which are systematically categorized into perception and reasoning domains. For in-distribution evaluation, the lite subset of MME-Realworld, consisting of 1,919 samples, is reserved as the test set, while the remaining 21,690 samples are utilized for training.

**V\* Bench.** [41] V\* Bench serves as an out-of-distribution benchmark, focuses on detailed visual grounding on high-resolution images. This vision-centric benchmark requires LMMs to accurately localize and interpret specific visual information, which has also been adopted by OpenAI to assess the visual reasoning capabilities of their latest o3 and o4-mini models [28]. This benchmark contains 191 test samples.

All datasets employ the multiple-choice question format, and model performance is consistently measured by accuracy on both the in-distribution (MME-Realworld) and out-of-distribution (V\* Bench) test sets. Figure 4 illustrates the distribution of image resolutions across different datasets.

### 4.2 Experimental Setup

We employ the verl [34] framework to enable distributed training across multiple machines and GPUs, and utilize vLLM [16] to accelerate inference during the rollout phase. For reinforcement learning, we adopt the naive GRPO [33] algorithm as RL baseline, where a post-prompt is added:

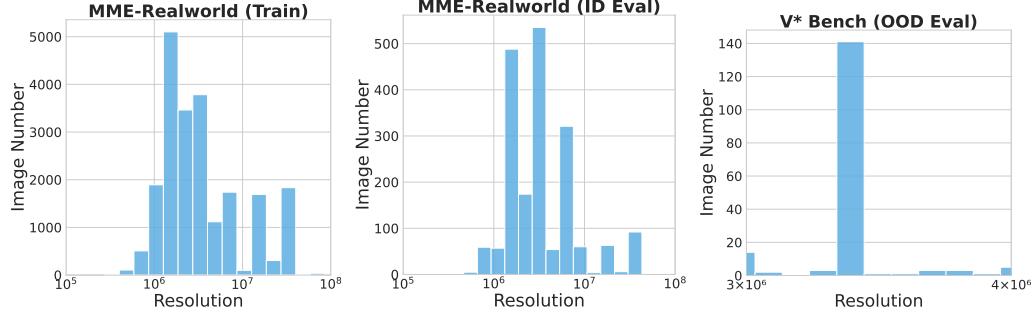


Figure 4: Distribution of image resolutions (width  $\times$  height) across different datasets.

*“{question}\nOutput the coordinates of the key image area relevant to the problem in JSON format. And put the answer letter (A, B, C, D, or E) within \boxed{}.”* Both GRPO and our proposed MGPO leverage a binary accuracy reward function, assigning a reward of 1 if the final multiple-choice answer is correct and 0 otherwise.

All experiments are conducted using the Qwen2.5-VL-7B [1] model. To prevent out-of-memory errors, the maximum number of input image pixels is limited to 1,003,520 ( $1280 \times 28 \times 28$ ), corresponding to a maximum of 1280 visual tokens per image. Images exceeding this pixel threshold are resized to comply with this constraint. More training details are provided in Appendix A.

### 4.3 Main Results

Table 1 presents the performance comparison of different post-training paradigms on Qwen2.5-VL-7B, including SFT, GRPO and our MGPO. All three post-training methods substantially improve the model’s performance on high-resolution visual tasks, as measured by both OOD V\* Bench and ID MME-Realworld benchmarks.

Notably, we observe that GRPO does not yield significant improvements over SFT, which contrasts with conclusions drawn from prior work on multi-modal mathematical tasks [26]. We hypothesize that, for high-resolution vision-centric tasks, the primary challenge lies in enabling the model to perceive fine-grained image details, rather than performing complex, lengthy reasoning.

In contrast, our MGPO algorithm achieves remarkable gains, outperforming both SFT and GRPO. Specifically, MGPO delivers a substantial 5.2% absolute improvement over the GRPO baseline on the V\* Bench (OOD) benchmark, and a 5.4% gain in overall MME-Realworld (ID) performance. These results demonstrate the effectiveness of multi-turn grounding and iterative sub-image cropping in addressing the challenges of high-resolution visual understanding.

Additionally, we compare our results with OpenAI’s o1 [14] and GPT-4o [12] models. To ensure a fair comparison, we mainly compare the OOD V\* Bench results. Notably, our MGPO post-trained model surpasses both o1 and GPT-4o, despite being based on a 7B model and trained with a small-scale dataset of 21k samples.

Figure 5 illustrates the comparative performance trajectories of MGPO and GRPO on the V\* Bench throughout the RL training process. As training progresses, MGPO consistently surpasses GRPO, highlighting its superior capacity to address high-resolution scenarios that remain unresolved by GRPO.

### 4.4 Emergent Grounding Ability During RL Training

In this section, we highlight the insight that it is feasible to train powerful grounding-based RL models even without grounding annotations. This insight can broaden the applicability of grounding-based RL paradigms, as obtaining high-quality grounding annotations is often expensive and labor-intensive.

To assess whether models can develop accurate grounding capabilities in the absence of grounding supervision, we analyze the proportion of rollouts that generate valid grounding coordinates during RL training (e.g., ensuring coordinates within the input image boundaries). Figure 6 illustrates the comparison between GRPO and MGPO. Regarding to GRPO, the ratio of valid grounding

Table 1: Performance comparison of different post-training paradigms for LMMs. V\* Bench serves as an out-of-distribution (OOD) evaluation, while MME-Realworld serves as an in-distribution (ID) evaluation. Abbreviations: OCR—Optical Character Recognition in the wild; RS—Remote Sensing; DT—Diagram and Table; MO—Video Monitoring; AD—Autonomous Driving.

Method	V* Bench (OOD)	MME-Realworld (ID)									
		Overall	Perception					Reasoning			
			OCR	RS	DT	MO	AD	OCR	DT	MO	AD
Qwen2.5-VL-7B	58.6	46.1	75.6	33.3	65.0	30.1	29.4	61.0	56.0	30.7	34.0
<i>+Post-Trianing (21K Samples):</i>											
SFT	71.7	58.7	85.6	<b>55.3</b>	78.0	43.6	43.7	70.0	60.0	50.7	<b>41.0</b>
GRPO	71.2	55.1	81.6	51.3	75.0	42.9	43.7	67.0	53.0	43.3	38.3
MGPO	<b>76.4</b> +5.2	<b>60.5</b> +5.4	<b>86.4</b>	54.0	<b>78.0</b>	<b>46.7</b>	<b>44.0</b>	<b>74.0</b>	<b>69.0</b>	<b>52.7</b>	39.3
<i>OpenAI’s Models:</i>											
OpenAI o1	69.7	-	-	-	-	-	-	-	-	-	-
OpenAI GPT-4o	73.9	45.2	77.7	28.9	46.7	33.9	22.4	61.4	44.8	36.5	26.4

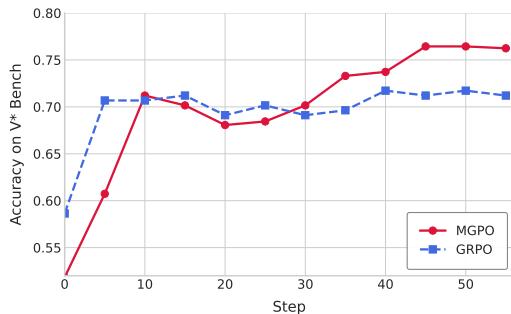


Figure 5: Performance comparison of V\* Bench between MGPO and GRPO.

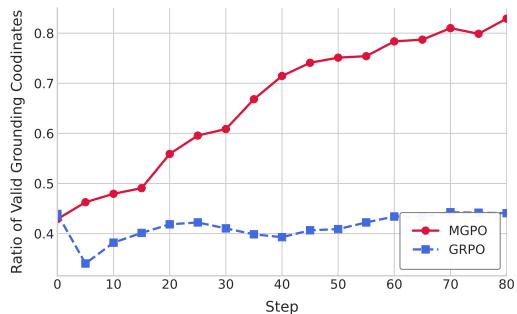


Figure 6: The ratio of valid grounding coordinates during RL rollouts.

coordinates remains low and exhibits minimal improvement throughout training, indicating that the model struggles to ground correct image regions. In contrast, MGPO demonstrates a clear upward trajectory, with the proportion of valid grounding coordinates steadily increasing as training progresses.

Additionally, we evaluate whether the grounding sub-images from the test set can be directly used to answer the question using Qwen2.5-VL-7B. As presented in Table 2, the comparative results across different methods demonstrate the superior accuracy of grounding achieved by MGPO. In the second stage of MGPO, the model is provided with either the cropped subimage or the original image, without any auxiliary reward for generating valid sub-image coordinates. Notably, the model autonomously increases the proportion of valid grounding coordinates, suggesting that it is capable of learning to localize key regions and utilize subimages to improve question answering performance.

We further conduct supplementary experiments on the image counting task in Appendix B, leveraging the fact that the image count dataset provides both the grounding annotations (in point format) and the corresponding count as the final answer. We compare two reward function for RL post-trianing: (1) the binary accuracy reward based solely on the correctness of the final count, and (2) incorporating an additional point reward based on grounding annotations. Both qualitative and quantitative results indicate that introducing the additional point reward does not yield significant performance improvements.

**Effect of LMM Maximum Input Image Resolution.** Table 3 compares the impact of varying maximum input image resolutions for LMMs. We observe that MGPO yields greater performance improvements on the V\* Bench when the maximum input pixel limit is lower. This is because, when high-resolution images are aggressively resized, many tasks become more challenging to solve

Table 2: Ratio of grounding subimages that can directly answer the question using Qwen2.5-VL-7B on the V\* Bench.

Method	Ratio (%)
Qwen2.5-VL	47.1
+GRPO	50.2
+MGPO	65.2

Table 3: Performance comparisons of various post-training paradigms for LMMs under different maximum input image resolutions.

Method	Max Input Image Pixels			
	$640 \times 28 \times 28$		$1280 \times 28 \times 28$	
	V* Bench (OOD)	MME-Realworld (ID)	V* Bench (OOD)	MME-Realworld (ID)
Qwen2.5-VL	50.3	40.9	58.6	46.1
<i>+Post-Training (21K Samples)</i>				
SFT	66.5	55.8	71.7	58.7
GRPO	64.4	54.0	71.2	55.1
MGPO	73.3 <sub>+8.9</sub>	57.0 <sub>+3.0</sub>	76.4 <sub>+5.2</sub>	60.5 <sub>+5.4</sub>

directly. However, MGPO can first identify key regions and crop clearer sub-images from the original image, which enables more effective task completion under constrained resolution settings.

## 5 Limitation

All experiments of MGPO are conducted using a fixed two-turn template, rather than allowing the model to autonomously decide when to perform image cropping based on the input question, as illustrated in lastest OpenAI models such as o3 and o4-mini [28]. This limitation stems from our observation that Qwen2.5-VL [1], when directly subjected to RL post-training, struggles to generate grounding coordinates without explicit prompt guidance.

Nevertheless, we believe that our trained models can be leveraged to generate high-quality chain-of-thought (CoT) data for subsequent SFT. By adopting a multi-stage training strategy that combines SFT and RL, as in DeepSeek-R1 [9], may ultimately enable the model to autonomously decide when and how to perform grounding. We leave this direction for future work.

## 6 Related Work

### 6.1 Large Multimodal Model

Recent advancements in LMMs [1–3, 7, 13, 17, 22, 24, 38, 44, 45] have equipped the model with robust visual understanding capabilities, by leveraging high-quality instruction tuning data [7, 17, 18], enhanced model architecture [22, 40], and meticulously designed training pipelines [24, 38]. Despite these efforts, most LMMs concentrate common-resolution visual inputs. Although some research [10, 35, 48] addresses high-resolution image analysis, they often require high training costs or fall short in generalization due to insufficient data. In this work, we aim to develop LLMs capable of high-resolution reasoning with reduced training costs and minimal data requirements.

### 6.2 Visual-Centric Multi-modal Reasoning

Previous research primarily employs the Chain-of-Thought paradigm to either simplify complex visual information [23, 32, 37] or enhance reasoning abilities through instruction tuning [8, 19, 42]. Recently, inspired by DeepSeek-R1 [9], there has been increasing interest [11, 27, 29, 46, 47] in using RL to enhance the reasoning capabilities of LMMs. Nevertheless, they often overlook the crucial role of visual-centric reasoning in multimodal contexts. In this work, we focus on visual-centric multimodal reasoning, employing RL algorithms to address complex visual tasks.

## 7 Conclusion

This paper introduces MGPO, a multi-turn grounding-based reinforcement learning algorithm to solve high-resolution real-world scenarios. By relying solely on a binary reward signal based on the correctness of the final answer, MGPO enables models to emerge robust grounding abilities during the RL training process, without requiring any grounding annotations. We hope our insight can advance the progress of grounding-based visual reasoning models.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, ZhaoYang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.
- [4] A. T. Clark et al. How many megapixels is the human eye?, 2014. Accessed on May 7, 2025.
- [5] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. Human photoreceptor topography. *Journal of Comparative Neurology*, 292(4):497–523, 1990.
- [6] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.
- [7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.
- [8] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.
- [9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [10] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.
- [11] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [14] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

- [16] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
- [17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [18] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025.
- [19] Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 2024.
- [20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [21] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.
- [22] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024.
- [23] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.
- [24] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*, 2025.
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [26] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhui Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- [27] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhui Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- [28] OpenAI. Openai o3 and o4-mini system card. <https://openai.com/index/o3-o4-mini-system-card/>, 2024. Accessed: 2025-04-18.
- [29] Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, et al. Skywork r1v: pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.
- [30] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.
- [31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- [32] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- [33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [34] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.
- [35] Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. *arXiv preprint arXiv:2503.19903*, 2025.
- [36] J. D. Smith et al. Foveal cone density and visual acuity. *Vision Research*, 150:45–53, 2018.
- [37] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*, 2024.
- [38] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- [40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [41] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.
- [42] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- [43] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [44] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. In *European Conference on Computer Vision*, pages 20–38. Springer, 2024.
- [45] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*, 2025.
- [46] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [47] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

- [48] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.
- [49] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal lilm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

## A Training Details

Model training is conducted on a computational cluster comprising four nodes, each equipped with eight H100 GPUs. Optimization is performed using the AdamW [25] optimizer with a fixed learning rate of  $1 \times 10^{-6}$ . The global training batch size is set to 512, and a mini-batch size of 128 is used for each iteration. During rollout, eight responses are sampled per prompt. The Proximal Policy Optimization (PPO) clip ratio [31], which constrains the magnitude of policy updates and plays a role analogous to the learning rate in reinforcement learning, is set to the default value of 0.2.

## B Further Experiments on Image Counting Tasks

To further substantiate the insight that training powerful grounding-based RL models even without grounding annotations, we conduct additional experiments on the image counting task. Specifically, we randomly sample 3,000 instances from the Pixmo-Points [7] dataset for post-training, which provides both the grounding annotations (in point format) and the corresponding count as the final answer. Pixmo-Count [7] is used as the in-distribution (ID) evaluation benchmark, while FSC-147 [30] serves as the out-of-distribution (OOD) benchmark.

During GRPO post-training, the model is prompted to first grounding (point) each object in the image and subsequently provide the total count. We compare two reward function: (1) the binary accuracy reward based solely on the correctness of the final count, and (2) incorporating an additional point reward. The point reward is computed by matching the model’s predicted point list with the ground-truth point list using the Hungarian algorithm [15], such that a higher number of matched ratio results in a higher reward.

The results, summarized in Table 4, indicate that introducing the additional point reward does not yield significant performance improvements. We further visualize the outputs of the GRPO model trained solely with the accuracy reward (see Figure 7), and observe that the model is capable of accurately localizing object points even without explicit grounding supervision. These results support our conclusion that explicit grounding annotations are not necessary for effective RL-based learning, as the model inherently learns to perform precise grounding as a prerequisite for solving the counting task.

Table 4: Performance comparison of image count task. Additional point reward do not lead to significant performance improvements.

Method	FSC-147 (OOD)	Pixmo-Count (ID)
Qwen2.5-VL-7B	59.9	13.0
+Post-Trianing (3K Samples)		
SFT	72.7	24.1
GRPO (Accuracy Reward)	81.0	35.0
GRPO (Accuracy + Point Reward)	81.9	34.9



Figure 7: Visualization of point predictions from the GRPO model trained with only accuracy reward.

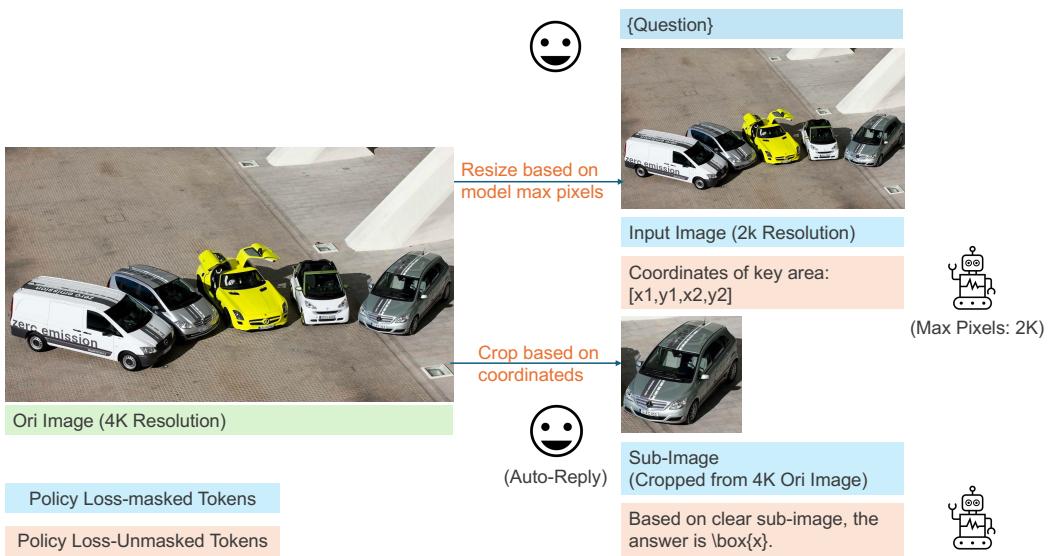


Figure 8: A illustration of cropping sub-image based on grounding coordinates.

```

<|im_start|>system
You are a helpful assistant. Given an image and one question. First, identify the coordinates
of the key image area relevant to solving the problem. Append the coordinates in JSON
format at the end of your response and stop. This will trigger cropping of the corresponding
area in the original image and enlarge it for improved clarity. Once the enlarged image is
available, provide the final answer (A, B, C, D, or E) enclosed within \boxed{}.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>
Question: What's the plate number of the grey car on the right side of this picture? Out the
coordinates of the key image area relevant to the problem in JSON format.<|im_end|>
<|im_start|>assistant
```json
[
    {"bbox_2d": [635, 260, 812, 424], "label": "key image area"}
]
```
<|im_end|>
<|im_start|>user
Based on the provided coordinates, return the sub-image:
<|vision_start|><|image_pad|><|vision_end|>
Question: What's the plate number of the grey car on the right side of this picture? The
choices are listed below:
(A) LB MA 864
(B) S EV 4201
(C) S FC 311
(D) S PC 311
(E) This image doesn't feature the plate number.
Carefully analyze both the original image and the enlarged sub-image to solve the question
step by step. If the sub-image does not provide sufficient information, refer to the original
image to find the solution. Present your reasoning clearly, and provide the final answer (A, B,
C, D, or E) enclosed within \boxed{}.<|im_end|>
<|im_start|>assistant
To solve this problem, we need to identify the key image area relevant to the question, which
is the grey car on the right side of the image. The coordinates provided in the JSON format
point to the correct area.

1. **Identify the key image area**: The coordinates provided in the JSON format indicate that
the key image area is the grey car on the right side of the image.
2. **Analyze the enlarged sub-image**: The enlarged sub-image shows the grey car on the
right side with a clear view of its license plate.
3. **Read the license plate number**: The license plate number on the grey car is clearly
visible as "S FC 311".

Based on the analysis, the plate number of the grey car on the right side of the picture is
\boxed{C}.<|im_end|>

```

Figure 9: A full conversation example of MGPO post-trained model on high-resolution image tasks.