

MGPO: Thinking with Images via Multi-Turn Grounding-Based Reinforcement Learning

Xinyu Huang^{1,2,*} Yuhao Dong^{2,*} Weiwei Tian¹ Bo Li² Rui Feng¹ Ziwei Liu^{2,†}

¹Fudan University ²S-Lab, Nanyang Technological University

*Equal contribution. †Corresponding author.

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. Codes are available at <https://github.com/EvolvingLMMs-Lab/MGPO>.

1 Introduction

State-of-the-art large multimodal model (LMM) architectures, such as Qwen2.5-VL (Bai et al., 2025), are typically built upon a powerful large language

model (LLM) backbone (e.g., Qwen2.5 (Yang et al., 2024)) integrated with an external Native Resolution Vision Transformer (NaViT) module (Dehghani et al., 2023; Tschannen et al., 2025). This module allows LMMs to process images at their original resolutions, enabling the model to capture fine-grained visual details and achieve high perceptual fidelity (Liu et al., 2024b; Li et al., 2024).

Nevertheless, this 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 (Liu et al., 2023). (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.

Inspired by the human visual system, although its theoretical resolution is estimated to be about 576 megapixels (Clark et al., 2014; Curcio et al., 1990), sharp vision is limited to the foveal area of the macula, which covers only around 125,000 pixels within the central 1° of the visual field (Clark et al., 2014; Smith et al., 2018). Therefore, when viewing high-resolution real-world scenes, the human visual system adopts task-driven visual search strategies (Wu and Xie, 2024) to locate and analyze important regions. Following this biological principle, we aim to give large multimodal models (LMMs) a similar visual search ability by using visual grounding to focus on key image areas.

However, enabling LMMs to perform grounding-based visual reasoning remains challenging, mainly because acquiring accurate grounding annotations for standard visual question answering (VQA) datasets is both scarce and costly. These annotations are essential for constructing multi-turn grounding conversations used in supervised fine-tuning (SFT). In this paper, we demonstrate that

accurate grounding behavior can emerge within a reinforcement learning framework, even when the training signal is limited to a binary reward derived solely from the correctness of the final answer.

To this end, we introduce **Multi-turn Grounding-based Policy Optimization (MGPO)**, a reinforcement learning (RL) framework built upon the Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) algorithm with a novel multi-turn grounding-based rollout strategy. MGPO 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. 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 endows LMMs with a top-down, question-guided visual search mechanism tailored for high-resolution scenarios, enabling interpretable reasoning by indicating the image regions that should be attended to throughout the inference process.
- **Overcomes Maximum Pixel Constraints.** MGPO overcomes the maximum pixel limitation of LMMs. Even when a high-resolution image must be resized to fit within pixel constraints, resulting in a blurred input, the model can still identify the relevant coordinates and extract clear sub-images from the original image for further analysis.

- **Without Additional Grounding Annotations.** MGPO can be post-trained directly on standard VQA datasets without any extra grounding annotations. In contrast to closely related work such as DeepEyes (Zheng et al., 2025), which still relies on ground-truth bounding boxes for data filtering and on intermediate grounding rewards during RL, MGPO is trained with the final-answer reward alone, yet still brings substantial improvements in intermediate grounding performance over GRPO (Shao et al., 2024b).

Ultimately, we use MGPO to post-train Qwen2.5-VL-7B (Bai et al., 2025) on visual-question-short answering data, and the resulting model achieves strong intermediate grounding performance without requiring grounding annotations. Compared to GRPO, MGPO yields a 5.4% improvement on the in-distribution MME-Realworld (Zhang et al., 2024b) benchmark and a 5.2% gain on the challenging out-of-distribution (OOD) V* Bench (Wu and Xie, 2024). Notably, leveraging with only 21K post-training samples, our model surpasses OpenAI’s o1 and GPT-4o models on the OOD V* Bench.

2 Related Work

2.1 Large Multimodal Models

Recent advancements in LMMs (Bai et al., 2025; Li et al., 2024; Liu et al., 2024c, 2025; Deitke et al., 2024; Hurst et al., 2024a; Team et al., 2025; Chen et al., 2024a,b) have equipped the model with robust visual understanding capabilities, spanning from simple tasks such as common image comprehension to more challenging ones like long video analysis and mathematical reasoning. By leveraging high-quality instruction tuning data (Li et al., 2024, 2025; Deitke et al., 2024), enhanced model architecture (Liu et al., 2024c; Wang et al., 2024), and meticulously designed training pipelines (Liu et al., 2025; Team et al., 2025), these models achieve fine-grained multi-modal alignment across various tasks. Despite these efforts, most LMMs concentrate on general visual understanding with common-resolution inputs, overlooking the inherent human ability to perceive fine-grained details in high-resolution images. Although some research (Zhang et al., 2024a; Guo et al., 2024; Shi et al., 2025) addresses high-resolution image analysis, it often falls short in generalization and competitive performance due to high training costs and

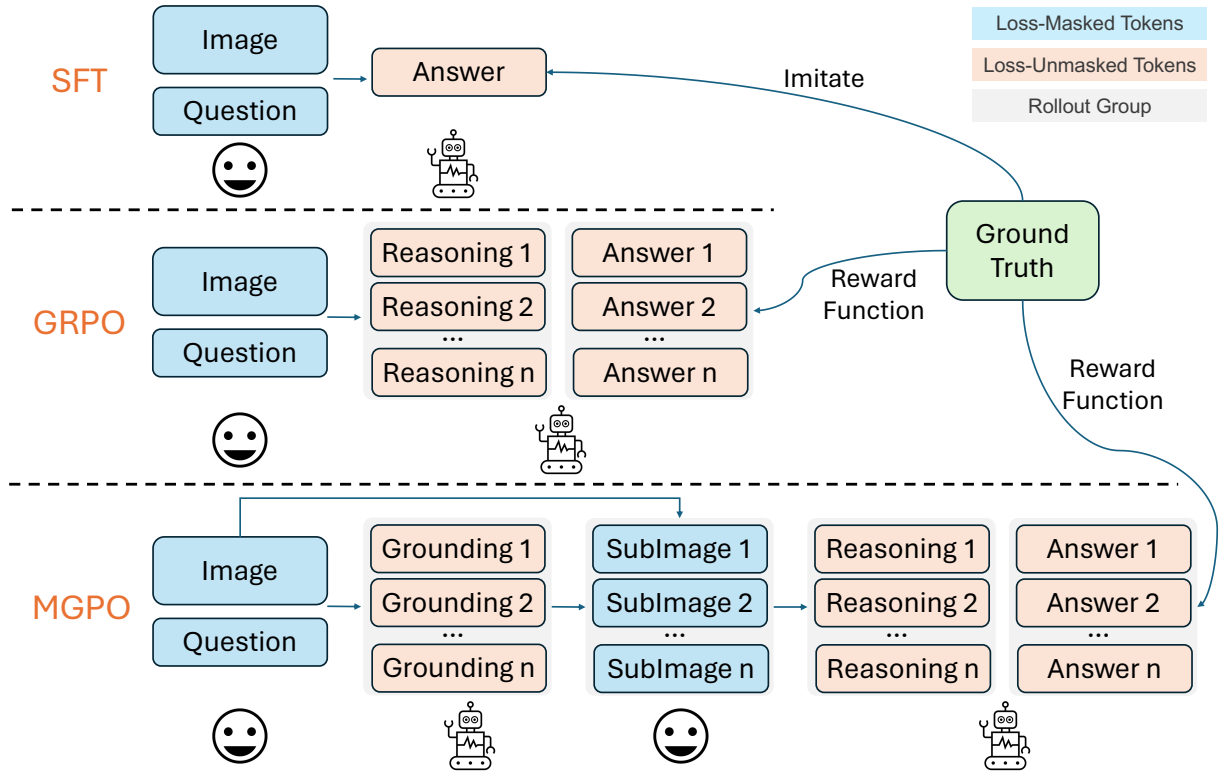


Figure 1: Comparison of different post-training paradigms for LMMs. Our MGPO automatically crops and returns sub-image to the model based on its predicted grounding coordinates, enabling the model to iteratively focus on key regions and effectively solve high-resolution visual tasks.

insufficient data. In this work, we aim to tackle these challenges by developing LMMs capable of high-resolution image understanding with reduced training costs and minimal data requirements.

2.2 Visual-Centric Multi-modal Reasoning

Multi-modal reasoning is a fundamental capability for large multimodal models (LMMs) when tackling complex and challenging tasks. Prior research has predominantly adopted the Chain-of-Thought (CoT) paradigm to either decompose intricate visual information into interpretable steps (Liu et al., 2024d; Shao et al., 2024a; Sun et al., 2024) or enhance reasoning performance through instruction tuning (Xu et al., 2024; Dong et al., 2024; Liu et al., 2024a). More recently, inspired by the success of DeepSeek-R1 (Guo et al., 2025), a growing body of work (Meng et al., 2025a; Huang et al., 2025; Yang et al., 2025; Zhang et al., 2025; Peng et al., 2025) has introduced reinforcement learning (RL) to further strengthen the reasoning abilities of LMMs. While these approaches substantially improve general reasoning across diverse tasks, they often underemphasize the ability to reason directly over visual content. Consequently, models may

perform well on mathematical problem-solving, yet still exhibit weaknesses in perception-oriented tasks that require accurate visual grounding. A closely related line of work, exemplified by DeepEyes (Zheng et al., 2025), explores “thinking with images” through RL, where the model iteratively grounds and crops sub-images to support reasoning. However, such methods still rely on ground-truth bounding boxes for data selection and on intermediate grounding rewards during RL. In this work, we address the same gap from a different angle: we develop a reinforcement learning framework that elicits accurate grounding behavior from the final-answer reward alone, removing the need for any grounding annotations.

3 Preliminaries

3.1 Advanced Baseline LMMs

We adopt Qwen2.5-VL (Bai et al., 2025) as our baseline, which combines the Qwen2.5 (Yang et al., 2024) language model with a Native Resolution Vision Transformer (NaViT) (Dehghani et al., 2023; Tschannen et al., 2025) that processes images at their native resolutions. NaViT divides an input image I into non-overlapping patches and aggre-

```

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 2: Fixed multi-turn grounding template employed during MGPO training. This template guides the model to first identify the relevant regions and then perform reasoning to produce the final answer. When integrated with the MGPO training pipeline, this approach enables the model to develop grounded reasoning capabilities without a separate cold-start SFT stage. The bounding-box snippet in the upper-right corner is a raw output from the early stage of training, where the model has not yet learned to emit coordinates only; the format becomes increasingly clean as RL proceeds (see Figure 5).

gates groups of $m \times m$ patches via an MLP into visual tokens X_i , which are concatenated with textual tokens X_t and jointly encoded for multi-modal reasoning. Qwen2.5-VL also supports visual grounding by predicting bounding box coordinates $[x_1, y_1, x_2, y_2]$, though grounding typically requires explicit prompts and is not inherently triggered during reasoning.

3.2 Single-Turn Multi-modal RL

Figure 1 compares post-training paradigms for LMMs. In supervised fine-tuning (SFT), the model directly imitates ground-truth responses, whereas reinforcement learning (RL) encourages exploration of reasoning trajectories that yield correct answers, going beyond the limits of imitation learning. We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) as our RL baseline. Given visual tokens X_i and textual context X_t , the model produces G candidate outputs $\{X_a^g\}_{g=1}^G$. A rule-based reward function \mathcal{R} assigns a binary score $r^g \in \{0, 1\}$ indicating whether each prediction is correct. The policy gradient is:

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{X_a^g \sim p_{\theta}} \left[(r^g - b) \cdot \nabla_{\theta} \log p_{\theta}(X_a^g | X_i, X_t) \right], \quad (1)$$

where $b = \frac{1}{G} \sum_{g=1}^G r^g$ is the average reward within the group. Only output tokens X_a^g contribute to the loss, while input tokens serve solely as conditioning context.

4 Multi-turn Grounding-Based RL

Motivation. While LMMs exhibit strong reasoning ability, they often fail to focus on the correct visual regions, which leads to inaccurate or hallucinated answers. When the model accurately localizes the regions referred to by the question, its reasoning becomes more coherent and the final answer more reliable; otherwise, misaligned attention frequently produces inconsistent reasoning steps. This observation motivates our multi-turn grounding-based RL strategy, which encourages the model to first attend to question-relevant regions before producing an answer, so as to better couple visual perception with reasoning in high-resolution scenarios.

4.1 Formulation

To address the challenges of high-resolution visual reasoning, we propose Multi-turn Grounding-based Policy Optimization (MGPO). *MGPO is not a new RL optimizer; rather, it is an instantiation of GRPO*

with a multi-turn grounding-based rollout, where sub-image cropping is interleaved with the model’s generation. Figure 1 shows the two-turn illustration compared with previous methods, and Algorithm 1 gives the general formulation. In this paradigm, the model operates over K sequential interactions, 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.

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:

$$\nabla_{\theta} \mathcal{J}_{\text{MGPO}}(\theta) = \mathbb{E}_{\{X_a^{(j),g}\}_{j \sim p_{\theta}}} \left[(r^g - b) \cdot \sum_{j=1}^{k_g} \nabla_{\theta} \log p_{\theta} \left(X_a^{(j),g} \mid \mathcal{H}^{(j)} \right) \right], \quad (2)$$

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, which ultimately contribute to the correct final answer.

4.2 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 hinders effective multi-turn RL. To address this, we design a fixed two-turn dialogue template, as shown in Figure 2, to explicitly activate the model’s grounding and reasoning abilities during the training process.

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

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

Input: Policy model π_{θ} ; 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 $X_a^{(k),g} \sim \pi_{\theta}(\cdot \mid \mathcal{H}^{(k)})$;

if $X_a^{(k),g}$ *contains final answer* **then**
└ **break**

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

$X_i^{(k+1)} \leftarrow$

Crop($X_i^{(k)}$, 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 π_{θ} using policy gradients calculated by Eq. 2

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 eliminates the cold-start stage, which would otherwise require constructing multi-turn grounding data and performing 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 design 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-images 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

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

$$[\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2] = \frac{[x_1, y_1, x_2, y_2]}{S_{\text{input}}} \cdot S_{\text{ori}}, \quad (3)$$

where S_{input} and S_{ori} represent the width and height of the input and original images, respectively. A detailed illustration of this process is provided in the Appendix Figure 9.

5 Experiments

5.1 Datasets & Metrics

To evaluate the effectiveness of our approach, experiments are conducted on two established datasets: MME-Realworld (Zhang et al., 2024b) and V* Bench (Wu and Xie, 2024). 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. (Zhang et al., 2024b) 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. (Wu and Xie, 2024) V* Bench serves as an out-of-distribution benchmark that 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 (OpenAI, 2024). 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 3 illustrates the distribution of image resolutions across different datasets.

5.2 Experimental Setup

We employ the verl (Sheng et al., 2024) framework to enable distributed training across multiple machines and GPUs, and utilize vLLM (Kwon et al., 2023) to accelerate inference during the rollout phase. For reinforcement learning, we adopt the

naive GRPO (Shao et al., 2024b) algorithm as RL baseline, where a post-prompt is added: “*{question}*”
Output 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 (Bai et al., 2025) 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 of MGPO are provided in Appendix A.

5.3 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 (Meng et al., 2025b). This abnormal result indicates 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 framework demonstrates significant performance gains, consistently outperforming both SFT and GRPO across various benchmarks. Specifically, MGPO achieves a 5.2% absolute improvement over the GRPO baseline on the V* Bench (OOD) benchmark and a 5.4% increase in overall performance on the MME-Realworld (ID) dataset. These improvements underscore the effectiveness of the proposed multi-turn grounding and iterative sub-image cropping strategies in enhancing the model’s high-resolution visual understanding and reasoning abilities.

To further assess MGPO’s competitiveness, we compare our results with OpenAI’s o1 (Jaech et al., 2024) and GPT-4o (Hurst et al., 2024b) models. For fairness, the comparison primarily focuses on

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-Training (21K Samples):</i>											
SFT	71.7	58.7	85.6	55.3	78.0	43.6	43.7	70.0	60.0	50.7	41.0
GRPO	71.2	55.1	81.6	51.3	75.0	42.9	43.7	67.0	53.0	43.3	38.3
MGPO	76.4 ^{+5.2}	60.5 ^{+5.4}	86.4	54.0	78.0	46.7	44.0	74.0	69.0	52.7	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

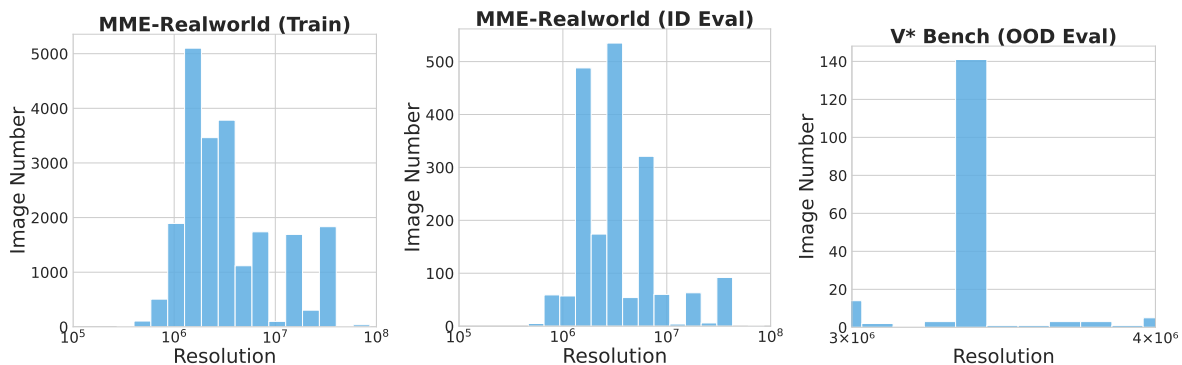


Figure 3: Distribution of image resolutions (width × height) across different datasets.

the V* Bench (OOD) results. Remarkably, despite being built upon a relatively lightweight 7B model and trained on a modest dataset containing only 21k samples, our MGPO-post-trained model surpasses both o1 and GPT-4o in fine-grained visual perception. This highlights MGPO’s high data efficiency and strong generalization capacity under limited training resources.

Figure 4 further illustrates the comparative training trajectories of MGPO and GRPO on the V* Bench. Although MGPO exhibits lower accuracy at the initial stage, which is mainly due to the imperfect formulation of grounding information during early training, it quickly surpasses GRPO as learning progresses. This improvement highlights MGPO’s superior capability in addressing complex high-resolution visual scenarios that remain challenging for GRPO.

Generalization across Model Sizes and Diverse Benchmarks.

To further examine how broadly

MGPO applies, we extend the evaluation in two directions: (i) a smaller base model, Qwen2.5-VL-3B, to check whether the gains carry over to a different model scale; and (ii) two additional benchmark families, HR-Bench (Wang et al., 2025) (4K and 8K) for high-resolution understanding and POPE (Li et al., 2023) for object hallucination. The results are summarized in Table 2. MGPO consistently outperforms the GRPO baseline on both 3B and 7B base models across all five benchmarks. The clear gains on HR-Bench suggest that multi-turn grounding effectively exploits fine-grained details in high-resolution images, while the improvements on POPE indicate that grounding answers in relevant visual evidence also helps reduce object hallucination.

Comparison with DeepEyes. DeepEyes (Zheng et al., 2025) is a closely related work that also explores “thinking with images” via reinforcement learning. Table 3 reports a head-to-head compari-

Table 2: Performance comparison on out-of-domain (V* Bench, HR-Bench), in-domain (MME-Realworld), and hallucination (POPE) benchmarks across two base model sizes.

Base Model	Method	V* Bench	MME-RW	HR-4K	HR-8K	POPE
Qwen2.5-VL-3B	GRPO	66.5	50.3	67.9	64.9	82.5
	MGPO	67.0	52.2	70.9	68.1	83.2
Qwen2.5-VL-7B	GRPO	71.2	55.1	69.8	66.3	84.5
	MGPO	76.4	60.5	74.2	71.2	86.1

Table 3: Comparison with DeepEyes on overlapping benchmarks. V* Bench is OOD for MGPO (no V* data used in training), whereas DeepEyes integrates V*-style data and uses grounding rewards with bounding-box supervision.

Base Model	Method	V* Bench	HR-4K	HR-8K
Qwen2.5-VL-7B	MGPO	76.4	74.2	71.2
	DeepEyes	91.3	75.1	72.6

son on overlapping benchmarks. Two differences in the training setup are worth noting when reading the absolute numbers: (i) DeepEyes includes V*-style data in training, whereas for MGPO V* is held out as an out-of-distribution benchmark and never seen during training; and (ii) DeepEyes uses additional reward signals beyond final-answer correctness (e.g., format and grounding rewards that rely on bounding-box supervision), while MGPO is trained only with the final-answer accuracy reward and uses no grounding annotations. Even under this weaker supervision, MGPO remains competitive on HR-Bench 4K/8K, which is consistent with our main claim that strong grounding behavior can emerge from final-answer rewards alone.

5.4 RL Brings Emergent Grounding Ability

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 5 illustrates the comparison between GRPO and MGPO. Regarding to GRPO, the ratio of valid grounding coordinates remains low and exhibits minimal improvement throughout training, indicating that the model

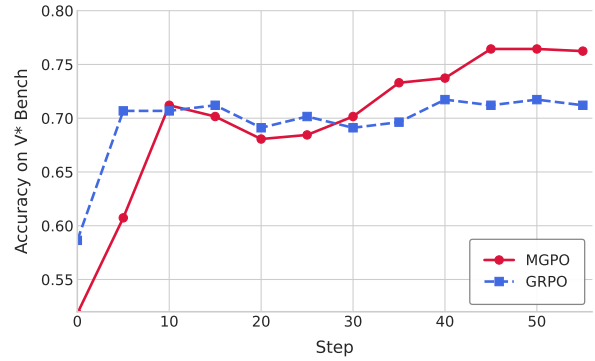


Figure 4: Performance comparison of V* Bench between our method (MGPO) and GRPO.

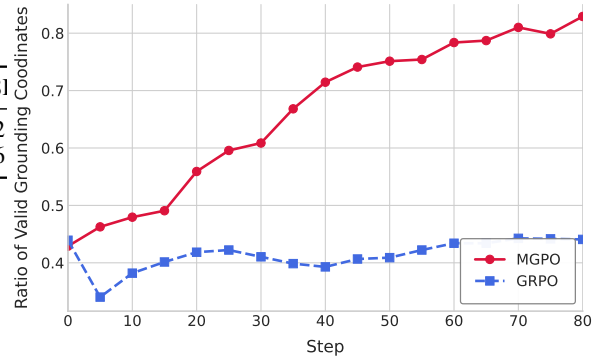


Figure 5: The ratio of valid grounding coordinates generated by models during RL rollouts.

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 grounded sub-images from the test set can be directly used to answer the question. As presented in Table 5, a significantly higher proportion (65.2%) of the sub-images generated by MGPO are sufficient for correctly answering the question, compared with 47.1% for the original Qwen2.5-VL and 50.2% for the GRPO baseline. This indicates that MGPO effectively identifies highly informative regions, even though no auxiliary reward is provided for generating valid sub-image coordinates: the model autonomously learns to localize key regions and utilize sub-images to improve question answering.

Robustness to Imperfect Grounding. Although MGPO consistently improves grounding quality, its final-answer correctness does not critically depend on perfect grounding at every step. The two-turn template provides a built-in safety net: in the answering turn the model still has access to the original global image from the first turn, so when a cropped sub-image is uninformative or inaccurate, it can fall back on the global context to produce an

Table 4: Performance comparison of various post-training paradigms for LLMs under different maximum input image resolutions. GRPO struggles with high-resolution image inputs, performing even worse than SFT as the resolution grows. This highlights the inherent difficulty of LLMs in handling fine-grained visual information. In contrast, MGPO achieves notable improvements by effectively identifying and focusing on critical image regions.

Method	Max Input Image Pixels			
	640 × 28 × 28		1280 × 28 × 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 ^{+8.9}	57.0 ^{+3.0}	76.4 ^{+5.2}	60.5 ^{+5.4}

Table 5: Ratio of grounded sub-images that can directly answer the question on V* Bench.

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

answer. In the worst case, MGPO is therefore no worse than the standard Qwen2.5-VL baseline that uses only the full image, while it benefits from focused sub-images whenever grounding is reliable.

We further conduct experiments on the image counting task in Appendix C, leveraging the fact that the image count dataset provides both the grounding annotations and the corresponding count as the final answer. We compare two reward functions for RL post-training: (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, supporting our design choice of relying solely on the final-answer reward.

Effect of Maximum Input Image Resolution. Table 4 compares the impact of varying maximum input image resolutions for LLMs. 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 directly. However, MGPO can first identify key regions and crop clearer sub-images from the original image, which enables more effective task completion un-

der constrained resolution settings.

Zero-Shot Generalization to More Interaction Turns. Although MGPO is trained with a fixed two-turn template for stability, the trained model is not tied to that interaction length at inference. We evaluate the trained 7B model on V* Bench by allowing one additional grounding round at test time, without any further training. As shown in Table 6, extending the interaction from 2 to 3 turns yields a further improvement (76.4 → 77.2), suggesting that MGPO has learned a transferable grounding-then-reasoning behavior rather than overfitting to the two-turn template.

Table 6: Zero-shot generalization to a 3-turn interaction at inference time on V* Bench (Qwen2.5-VL-7B + MGPO).

Evaluation Setting	V* Bench
2-turn (training setup)	76.4
3-turn (zero-shot at inference)	77.2

6 Conclusion

This paper introduces MGPO, a multi-turn grounding-based reinforcement learning framework built on GRPO with a grounding-aware rollout strategy, which elicits a thinking-with-images behavior for high-resolution real-world scenarios. Using only a binary reward derived from the correctness of the final answer, MGPO enables the model to develop robust grounding abilities during RL training, without requiring any grounding annotations. We hope the insights from MGPO can further advance research on grounding-based visual reasoning and contribute to more generalizable vision-language models.

Limitations

All experiments of MGPO are conducted using a fixed two-turn template during training, rather than allowing the model to autonomously decide when to perform image cropping based on the input question, as illustrated in lasted OpenAI models such as o3 and o4-mini (OpenAI, 2024). This limitation stems from our observation that Qwen2.5-VL (Bai et al., 2025), when directly subjected to RL post-training, struggles to generate grounding coordinates without explicit prompt guidance. That said, the trained model is not strictly tied to the two-turn setting at inference: as reported in Table 6, it can zero-shot generalize to a 3-turn interaction and obtain a further improvement on V* Bench, suggesting that the learned grounding-then-reasoning behavior is not merely tied to the fixed template.

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 (Guo et al., 2025), may ultimately enable the model to autonomously decide when and how to perform grounding. We leave this direction for future work.

Ethical Considerations

This research adheres to the ethical standards and guidelines for responsible AI research. The proposed MGPO framework and all associated experiments are conducted using publicly available datasets that do not contain personally identifiable information or sensitive content. All dataset licenses are respected, and appropriate citation and use policies are followed.

The development of MGPO aims to advance grounding-based visual reasoning rather than enable harmful or deceptive uses of AI technologies. The authors recognize the potential social and ethical implications of LMMs, including possible biases inherited from training data and unintended misuse in sensitive domains. To mitigate these risks, evaluation and analysis procedures are transparently reported, and results are carefully interpreted within the intended research scope. The authors encourage future work to continue improving explainability and transparency in grounding-based reinforcement learning systems.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. 2024. *Lmms-eval: Accelerating the development of large multimodal models*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. 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.
- A. T. Clark and 1 others. 2014. How many megapixels is the human eye? Accessed on May 7, 2025.
- C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson. 1990. Human photoreceptor topography. *Journal of Comparative Neurology*, 292(4):497–523.
- Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, and 1 others. 2023. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.
- Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jing Kang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2024. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zhanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In

- European Conference on Computer Vision*, pages 390–406. Springer.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024a. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024b. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, and 1 others. 2025. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*.
- Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. 2024a. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024c. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.
- Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. 2024d. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*.
- Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *arXiv preprint arXiv:2502.04328*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025a. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025b. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.
- OpenAI. 2024. Openai o3 and o4-mini system card. <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: 2025-04-18.
- Yi Peng, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, and 1 others. 2025. Skywork r1v: pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*.
- Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. 2021. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024b. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and 1 others. 2025. Scaling vision pre-training to 4k resolution. *arXiv preprint arXiv:2503.19903*.
- J. D. Smith and 1 others. 2018. Foveal cone density and visual acuity. *Vision Research*, 150:45–53.
- Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. 2024. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. 2025. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Penghao Wu and Saining Xie. 2024. 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.
- Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.
- Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. 2024a. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*.
- Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, and 1 others. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.
- Ziwei Zheng, Michael Yan, Jiang Chen, Jinghan Wang, Zhenhao Lai, Lewei Mei, Yuwei Cheng, Sheng Yan, Yiwu Cao, Yang Liu, and 1 others. 2025. Deepeyes: Incentivizing “thinking with images” via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Supplementary

In the supplementary material, we first present additional implementation and training details of the proposed MGPO framework in Section A. After that, we include an LLM usage section in Section B, detailing how we use LLMs properly in this paper. Then, in Section C, we provide further insights into the optimization and behavioral characteristics of grounding-based reinforcement learning models through experiments on image counting tasks. Finally, Section D reports extended qualitative results of the MGPO framework, offering a more comprehensive understanding of its performance across various visual reasoning scenarios.

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 (Loshchilov and Hutter, 2017) optimizer with a fixed learning rate of 1×10^{-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 (Schulman et al., 2017), 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. We strictly adhere to the Apache-2.0 License of the training data from MME-RealWorld (Zhang et al., 2024b). We use LMMs-Eval (Bo Li* and Liu, 2024) to evaluate our model.

B LLM Usage

In this paper, we clarify that Large Language Models (LLMs) are employed solely to support and refine the writing process. Specifically, we use LLMs to provide sentence-level suggestions and to enhance the overall fluency of the text.

C 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 (Deitke et al., 2024) dataset for post-training, which provides both the grounding annotations (in point format) and the corresponding count

as the final answer. Pixmo-Count (Deitke et al., 2024) is used as the in-distribution (ID) evaluation benchmark, while FSC-147 (Ranjan et al., 2021) serves as the out-of-distribution (OOD) benchmark.

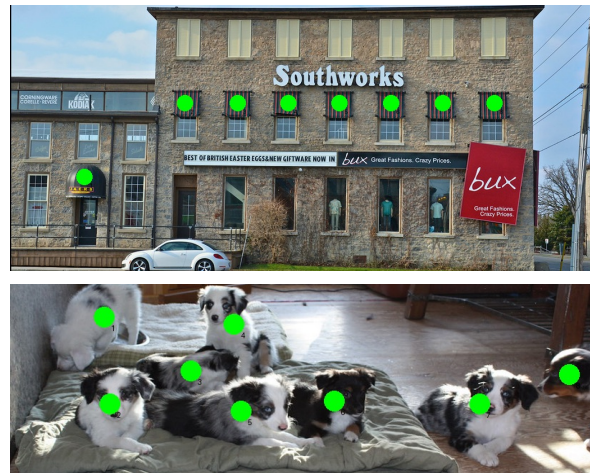


Figure 6: Visualization of point predictions from the GRPO model trained with only an accuracy reward.

Table 7: Performance comparison of the image count task. Additional point rewards do not lead to significant performance improvements.

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

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 (Kuhn, 1955), such that a higher ratio of match ratio results in a higher reward.

The results, summarized in Table 7, 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 accuracy reward (see Figure 6), 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 counting tasks.

D Qualitative Results

We provide more qualitative results to demonstrate the effectiveness of MGPO in Figure 7 and Figure 8. We visualize both the grounding results and final answer. MGPO enables the model to emerge with accurate grounding capabilities with only a binary reward, leading to the final correct answer.

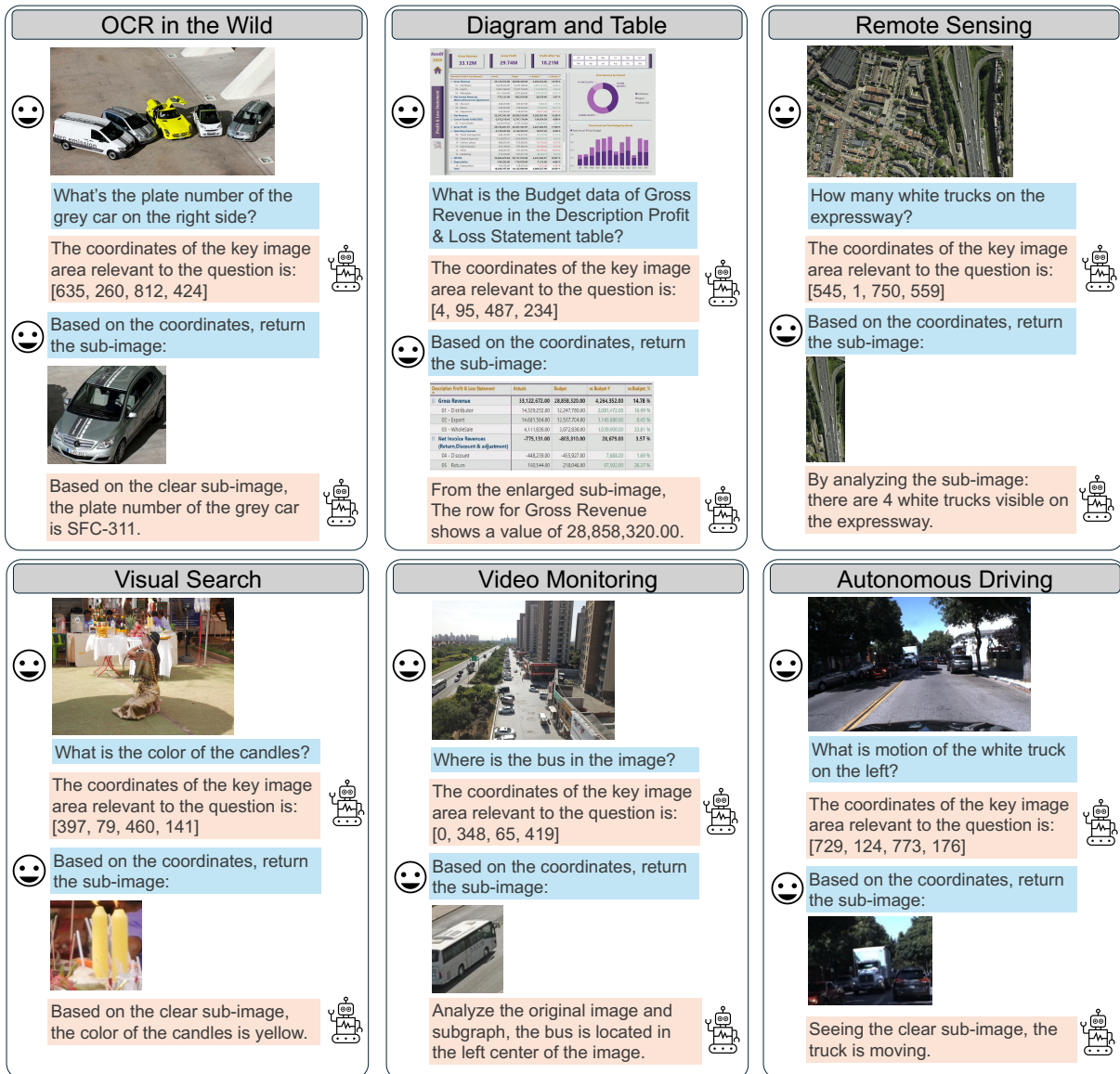


Figure 7: Examples of models trained with multi-turn grounding-based RL on high-resolution real-world tasks. The model first identifies key regions, which are then automatically cropped and returned as sub-images. Notably, despite only a binary reward function derived from the correctness of the final answer, the model gradually emerges with robust grounding capability throughout the RL process. The conversation in the figure only shows key parts, the full conversation is provided in Appendix Figure 8.

```

<|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 8: A full conversation example of the MGPO post-trained model on high-resolution image tasks.

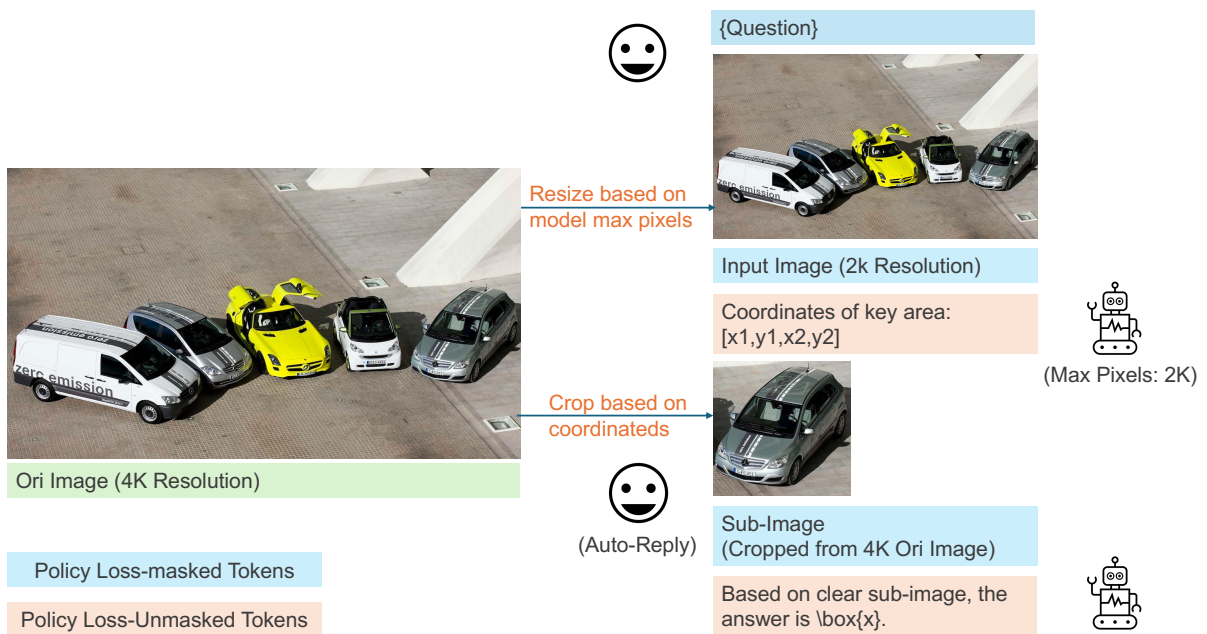


Figure 9: An illustration of cropping a sub-image based on grounding coordinates.