HART: High-Resolution Annotation-Free Reasoning Technique through a Closed-loop Framework

Jiacheng Yang[1*], Anqi Chen[1*], Yunkai Dang[1], Qi Fan[1], Cong Wang[1,2], Wenbin Li[1,3†], Feng Miao[2†], and Yang Gao[1]

[1] State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
[2] Institute of Brain-inspired Intelligence, Nanjing University, Nanjing, China
[3] Shenzhen Research Institute of Nanjing University, Shenzhen, China
liwenbin@nju.edu.cn, miao@nju.edu.cn

Abstract. Current Large Multimodal Models (LMMs) struggle with high-resolution visual inputs during the reasoning process, as the number of image tokens increases quadratically with resolution, introducing substantial redundancy and irrelevant information. A common practice is to identify key image regions and refer to their high-resolution counterparts during reasoning, typically trained with external visual supervision. However, such visual supervision cues require costly grounding labels from human annotators. Meanwhile, it remains an open question how to enhance a model's grounding abilities to support reasoning without relying on additional annotations. In this paper, we propose ***H**igh-resolution **A**nnotation-free **R**easoning **T**echnique (HART)*, a closed-loop framework that enables LMMs to focus on and self-verify key regions of high-resolution visual inputs. HART incorporates a post-training paradigm in which we design *Advantage Preference Group Relative Policy Optimization (AP-GRPO)* to encourage accurate localization of key regions without external visual annotations. Notably, HART provides explainable reasoning pathways and enables efficient optimization of localization. Extensive experiments on MME-RealWorld-Lite, TreeBench, V* Bench, HR-Bench-4K/8K, and MMStar demonstrate that HART improves performance across a wide range of high-resolution visual tasks, consistently outperforming strong baselines. Code will be available at https://github.com/RL-MIND/HART.

1 Introduction

Recent advances in Large Multimodal Models (LMMs) have attracted increasing attention from both industry and academia [3, 6, 10, 11, 17, 18]. LMMs have been widely applied to complex real-world tasks, such as object detection [25, 27] and visual question answering [43, 50, 55]. They demonstrate strong capabilities

* Equal contribution.
† Corresponding authors.

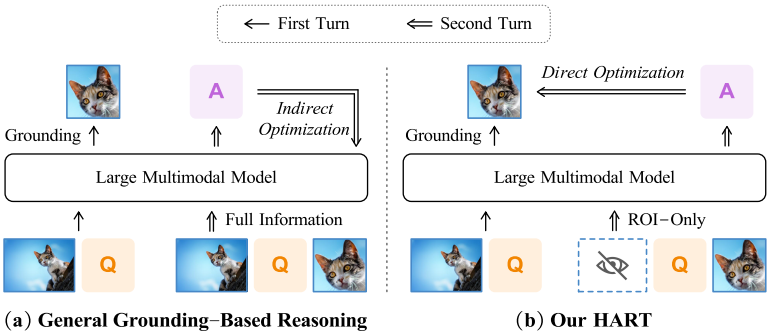

Fig. 1: Optimization procedures of (a) general grounding-based methods without bounding-box annotations and (b) our proposed model. General models indirectly optimize grounding performance, while HART performs direct optimization by answering based solely on the ROIs. Abbreviations: Q—Question; A—Answer.

in visual understanding and have the potential for enabling image-text interleaved reasoning. Despite these advances, current LMMs still face a critical limitation: their performance degrades significantly on challenging high-resolution visual tasks [15, 21, 60]. In such tasks, the number of visual tokens increases significantly with the resolution of the input images, while only a small subset contains key information. Popular LMM architectures, such as Qwen2.5-VL [3] and InternVL3 [68], typically impose a maximum pixel constraint on input images to address this issue. However, this constraint can lead to the loss of key information, restricting the model's visual perception capability [2, 7, 28, 64].

To address this issue, existing works have explored a reasoning pathway that incorporates visual grounding, which is inspired by human visual processing [29, 52, 60]. Humans need to identify food and predators for survival in the wild, thus evolving the macula, a zone of acute vision in the retina [38]. This structure guides visual attention and eye movements to locate key regions within high-resolution images. Previous works have developed visual grounded reasoning models, attempting to equip LMMs with similar structures and functions [31, 66]. Conditioned on the textual question, they first predict key regions of interest (ROI) with the downsampled image and then solve the question based on both the downsampled image and the ROIs from the original resolution. This pathway focuses only on critical visual information, thereby effectively reducing redundant computations. There are two lines of research that study the optimization of visual grounded reasoning models. On the one hand, some works directly enhance localization capabilities by using auxiliary visual annotations [34, 39, 41]. However, these methods require costly grounding labels from human annotators [48]. On the other hand, recent research leverages reinforcement learning (RL) [44] to jointly optimize grounding and reasoning without relying on additional annotated data [16, 42, 67]. These annotation-free approaches perform simple answer matching through a reward function. The model's rewards measure the correctness of the final answer but cannot directly reflect localization

accuracy. Specifically, the model receives a positive reward when the answer is correct, even if the localization is incorrect. Such reward misspecification will lead to negative optimization of grounding performance. In early experiments, we found this issue to be relatively common, occurring in 36.5% of cases for Qwen2.5-VL-7B [3] and 63.8% for InternVL3-8B [68]. Therefore, it is natural to think: ***How to directly optimize the grounding capabilities of LMMs without external visual annotations?***

To bridge this gap, this work aims to enable the model to self-verify its localization results for policy updates. We propose a novel approach, ***H**igh-resolution **A**nnotation-free **R**easoning **T**echnique (HART)*, to efficiently improve the performance of LMMs in high-resolution real-world scenarios without relying on extra visual annotations beyond final answers. We design a closed-loop framework to overcome the limitation imposed by resolution constraints. Given a textual question and a high-resolution image, our model identifies the ROIs and then crops relevant regions. These regions serve as visual feedback guiding the reasoning process. Subsequently, the original image is deliberately withheld, and the model answers the same question based solely on the cropped sub-images. Fig. 1 compares the procedures of general annotation-free grounding-based methods and our proposed method. Through this feedback loop, we introduce *Advantage Preference Group Relative Policy Optimization (AP-GRPO)*, a reinforcement fine-tuning strategy that alleviates the reward misspecification problem and directly enhances grounding capabilities by introducing dynamic hyper-parameter adjustment. Different from existing models [16, 67], HART can directly optimize visual grounding, thereby further enhancing the model's performance on perception-heavy tasks. When applied to post-train Qwen2.5-VL-7B [3], HART achieves significant improvements across a range of challenging high-resolution benchmarks, *i.e.*, +20.1% on MME-RealWorld [65], +6.7% on TreeBench [48], +2.1% on V* Bench [54], and +10.9% on HR-Bench-8K [51]. The main contributions are summarized as follows:

- We develop HART, a novel and interpretable framework that enhances the joint understanding of visual and textual inputs. It enables direct optimization of visual grounding without additional manual annotations.
- We introduce a reinforcement fine-tuning strategy, termed AP-GRPO, within the post-training paradigm to better incentivize the model to focus on key regions by prioritizing samples with correct grounding.
- We validate our method's effectiveness on several high-resolution visual benchmarks and show that HART achieves state-of-the-art performance among methods supervised only by the final answer.

2 Related Work

2.1 Large Multimodal Models

Recent breakthroughs in LMMs have significantly enhanced visual understanding capabilities, leading to growing popularity and widespread application across

various domains [20, 24, 37, 57, 58, 62, 64]. Inspired by DeepSeek-R1 [11], many of these methods leverage reinforcement learning (RL) with reward engineering to shape effective thinking processes and solve increasingly complex reasoning tasks [12, 32, 45, 47, 49, 56, 61, 62]. Despite recent progress, most LMMs still struggle with high-resolution image inputs and typically impose a resolution constraint due to limited visual-token capacity [13, 23, 26, 28]. This can result in blurred images and loss of key visual information, resulting in performance degradation. Such limitations further restrict the application of LMMs in high-resolution real-world scenarios, such as remote sensing and autonomous driving [7]. While recent works [14, 63] attempt to address this issue, they often overlook the crucial role of visual grounding in multimodal reasoning. In contrast, our work adopts a visual grounded reasoning process to address the challenges posed by high-resolution image analysis.

2.2 Visual Grounded Reasoning Models

Visual grounding LMMs aim to predict and focus on the key ROIs before actually answering the question [52]. Existing studies highlight that this process is similar to human visual processing and can ensure more accurate answers through grounded reasoning [29, 31, 60, 66]. There are two lines of work that study the optimization of visual grounding capabilities. One direction focuses on directly fine-tuning models on labeled data that contain ground-truth bounding box annotations of key image regions [34, 39, 41]. The differences between the predicted and ground-truth bounding boxes are used to update the policies. While such approaches can directly optimize the grounding capabilities of LMMs, they require large-scale and costly manual annotations [22].

Another line of research seeks to leverage end-to-end RL to improve grounding capabilities without external visual annotations [16, 42, 67]. However, these RL-based approaches typically post-train the model leveraging a reward signal derived from only the correctness of the final answer [48]. Consequently, they do not directly optimize grounding performance. We observed that annotation-free methods that directly optimize grounding capabilities remain undiscovered, as it is challenging to assess the grounding accuracy without bounding box annotations. Leveraging visual feedback, we aim to directly optimize the grounding capability of LMMs. Therefore, our work provides a more flexible solution that enables the model to self-verify its localization results for policy updates.

3 Preliminaries

3.1 Grounding-Based Visual Reasoning

Visual grounding refers to the task of localizing visual elements in an image based on a linguistic question. Grounding capabilities enable existing models to focus on critical visual regions that correspond to the given linguistic concepts, thereby effectively reducing redundant computations in high-resolution

vision-centric tasks [29, 66]. Formally, given an input image I_f and a textual question q, the visual grounding model first identifies the key sub-images I_s that semantically align with the entities or relations described in q, denoted as $I_s \sim \pi_\theta(\cdot \mid I_f, q)$, where π_θ denotes the model policy parameterized by θ. Next, the model leverages the task-relevant regions and generates an answer a based on both the full image and key sub-images, that is, $a \sim \pi_\theta(\cdot \mid I_f, I_s, q)$.

3.2 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) [40] is an efficient RL algorithm for post-training. For a question-answer pair, GRPO first generates a group of G candidate responses $\{o_i\}_{i=1}^G$ and receives corresponding rewards $\{r_i\}_{i=1}^G$. The advantage A_i of each response is computed in a group-relative manner:

$$A_i = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^G)}{\text{std}(\{r_i\}_{i=1}^G)}, \tag{1}$$

where $\text{mean}(\cdot)$ and $\text{std}(\cdot)$ are the mean and standard deviation of the rewards. The policy π_θ is updated as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right), \tag{2}$$

where β balances the KL-penalty term, $\pi_{\theta_{\text{old}}}$ is the old model, and π_{ref} is the reference model.

4 The Proposed Method

In this section, we begin with pilot experiments designed to illustrate the challenges faced by annotation-free visual grounding methods, which serve as the research foundation for our study. Then, we introduce the idea of our approach HART, a closed-loop framework that directly optimizes grounding and reasoning without relying on additional annotated data.

4.1 Vanishing Advantages in Indirect Optimization

The rewards of annotation-free visual grounding methods measure the correctness of the final answer but cannot directly reflect grounding accuracy. As a result, they receive a positive reward when the final answer is correct even though the grounding is incorrect, as illustrated in Fig. 2(a). This reward misspecification problem may lead to performance limitations. While this phenomenon is theoretically possible, its real-world occurrence and frequency remain unclear. Therefore, we conduct pilot experiments on LMMs grounding and reasoning, taking Qwen2.5-VL-7B [3] and InternVL3-8B [68] as representative examples.

Table 1: The joint distribution of answer correctness and grounding correctness for Qwen2.5-VL-7B [3] and InternVL3-8B [68] on the Visual CoT dataset [39].

Method	HART (Ours)	Correct answer		Incorrect answer		Proportion of incorrect grounding given correct answer
		Incorrect grounding	Correct grounding	Incorrect grounding	Correct grounding	
Qwen2.5-VL-7B [3]	✗	1057	1838	466	681	36.5%
	✓	**502**	1830	1028	682	**21.5%**
InternVL3-8B [68]	✗	1359	770	1578	335	63.8%
	✓	**998**	788	1931	325	**55.9%**

Experimental settings. Both models are evaluated on the test set of the Visual CoT dataset [39], since it provides ground-truth bounding box annotations of key image regions. Visual CoT contains a series of visual tasks, including text/doc, fine-grained understanding, charts, and relation reasoning. We apply a filtering procedure to exclude (a) subsets that ask descriptive questions or have complex answer formats, to facilitate answer validation, and (b) questions with yes/no answers, to reduce the impact of random guessing on the statistics. The resulting test set contains $4,042$ questions in total, each with one corresponding ground-truth bounding box. We adopt intersection over ground-truth [33, 59] as the metric to evaluate grounding accuracy. Specifically, grounding is considered correct if at least one predicted bounding box covers more than 0.3 of the ground-truth area.

Experimental results. Tab. 1 shows the distribution of the models' responses by answer accuracy and grounding accuracy. The responses are divided into four categories: correct answer with incorrect grounding, correct answer with correct grounding, incorrect answer with incorrect grounding, and incorrect answer with correct grounding. Qwen2.5-VL-7B correctly answers $2,895$ questions on Visual CoT, of which $1,057$ bounding boxes are incorrectly localized. Another popular model, InternVL3-8B, correctly answers $2,129$ questions on Visual CoT, of which $1,359$ bounding boxes are incorrectly localized.

Conclusions of pilot experiments. Existing annotation-free visual grounding methods typically post-train LMMs using only the final answer as the training signal [16, 67]. Based on our pilot experiments, we identify two limitations of these methods. First, it becomes difficult to quantify the actual contribution of the grounding. Second, in more than 36.5% of the cases where the models receive a positive reward, the localization is incorrect. Such reward misspecification can lead to negative optimization of grounding performance, as the model policy tends to encourage unreliable reasoning process. These limitations further restrict performance on visually intensive tasks.

4.2 HART

To overcome the above limitations, a novel closed-loop framework named ***H**igh-resolution **A**nnotation-free **R**easoning **T**echnique (HART)* is proposed to directly optimize high-resolution visual grounding and understanding. As illustrated in Fig. 2(b), we extend the idea of decomposing the reasoning process

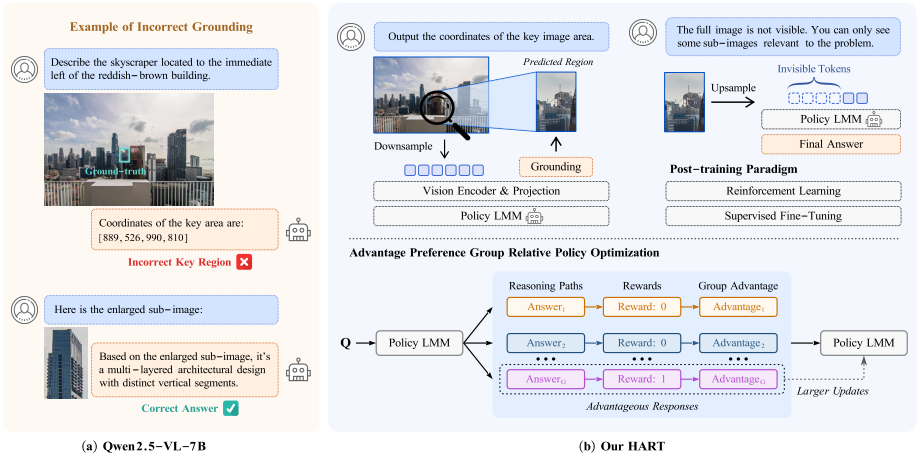

Fig. 2: Left: An example of Qwen2.5-VL-7B where the final answer is correct but the grounding is incorrect. Right: **HART Framework.** The post-training strategy consists of Reinforcement Learning (RL) and Supervised Fine-Tuning (SFT). In stage 1, after identifying the ROIs, the model answers based solely on the sub-regions and the original question. AP-GRPO is introduced to improve the model's grounding capabilities. In stage 2, HART uses SFT to further enhance the high-resolution reasoning capabilities.

from current research [22] to visual grounding LMMs. We also take into account the maximum-token constraint adopted in many models and restrict image resolution in the first turn. During the training phase, our proposed HART first prompts the model to predict the key ROIs conditioned on the down-sampled full input image I_f and the textual question q. The instruction is: *Output the coordinates of the key image area relevant to the problem* $[I_f, q]$. The model's response is returned in bounding-box format. In the second step, HART assesses whether the visual grounding is sufficient for generating the final answer. The ROIs I_s are cropped from the original high-resolution image, and the original image is deliberately withheld. The instruction is adjusted based on the visual feedback: *You were supposed to answer a question based on a full image, but now the full image is not visible. You can only see some sub-regions* $[I_s]$ *relevant to the problem. Answer the following question:* $[q]$. As shown in Tab. 1, through the above procedure, the proportion of Qwen2.5-VL-7B [3] and InternVL3-8B [68] producing a correct answer while localizing an incorrect region decreases by 15.0% and 7.9%, respectively. This result indicates that under the HART framework, the models' responses more accurately reflect the reliability of localization.

Let $L \in \{0, 1\}$ denote whether the model localizes the correct region and $R \in \{0, 1\}$ denote whether it produces the correct response. Compared to baseline pipelines [16, 67], the probability of our model producing a correct answer while localizing the incorrect region is much lower, leading to a smaller entropy. We state our result as follows.

Proposition 1. *Let $I_{HART}(L;R)$ and $I_{baseline}(L;R)$ denote the mutual information between localization correctness L and response correctness R under our method and baselines, respectively. Then,*

$$I_{HART}(L;R) > I_{baseline}(L;R). \tag{3}$$

We prove this proposition in Appendix A. The above expression shows that HART strengthens causal dependency between localization and reasoning. The feedback loop offers two benefits: (a) It enables the model to self-verify its localization results, eliminating the need for manual bounding box annotations. (b) The model can access more detailed visual information as we zoom in on the key region of the original image. This design reduces redundant computations and overcomes maximum pixel constraints imposed by existing LMMs, thereby enhancing the effectiveness in high-resolution real-world scenarios.

4.3 Advantage Preference Policy Optimization

We propose a post-training strategy to further improve the model's grounding and reasoning capabilities. In the HART framework, a correct final response is more likely to indicate that the localized ROIs contain all the visual information needed to answer the question. Hence, rewarding correct answers also encourages faithful grounding. Motivated by this observation, we proceed to modify the standard GRPO algorithm [40] to optimize the localization policy.

In vanilla GRPO, all samples are assigned equal weight. However, samples with incorrect responses clearly exhibit higher localization uncertainty, which can lead to reward misspecification within our feedback loop. Therefore, different from the traditional post-training methods, we prefer correct and advantageous responses as they typically indicate faithful grounding. Accordingly, we propose ***A**dvantage **P**reference **G**roup **R**elative **P**olicy **O**ptimization (AP-GRPO)* to assign dynamic weights to responses based on their advantages, thus allowing the model to focus more on optimizing the samples with correct grounding. In particular, given a textual question and the cropped sub-images, AP-GRPO first generates a group of candidate responses and receives corresponding rewards $\{r_i\}_{i=1}^{G}$, where $r_i = 1$ indicates a correct answer and $r_i = 0$ indicates an incorrect answer. Then the advantages of each response are computed, and the optimization objective is as follows:

$$\mathcal{J}_{\text{AP-GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}(\mu_1 \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i - \mu_2 \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})), \tag{4}$$

$$\mu_1 = 1 + k(r_i - \text{mean}(\{r_i\}_{i=1}^{G})), \tag{5}$$

$$\mu_2 = \beta \left(1 - k\big(r_i - \text{mean}(\{r_i\}_{i=1}^{G})\big)\right), \tag{6}$$

where the scaling factor k is the only hyperparameter in HART. First, the samples with correct grounding are assigned higher weights with the parameter μ_1, which provides larger updates for advantageous responses. Next, we introduce a

dynamic weighting for the KL penalty coefficient. The dynamic weight factor μ_2 reduces the KL penalty when the grounding is correct, thereby allowing greater deviation from the reference model.

Theoretical guarantees of AP-GRPO. To theoretically understand the advantages of the proposed AP-GRPO in comparison to prior studies, the following proposition verifies that AP-GRPO reduces reward misspecification that typically causes negative optimization of grounding performance. The proof is provided in Appendix B.

Proposition 2. *Let $g_{AP\text{-}GRPO}$ and $g_{baseline}$ denote the gradients of AP-GRPO and baselines, respectively. Consider any single-step reinforcement learning environment with binary rewards. Then $\exists \alpha \in [0, 1]$ such that*

$$g_{AP\text{-}GRPO} = g_{baseline} - \alpha P(L = 0, r = 1)\mathbb{E}_{L=0,r=1}\left[\nabla_\theta \log \pi_\theta\right]. \tag{7}$$

In other words, Proposition 2 demonstrates that AP-GRPO effectively reduces the negative impact of reward misspecification seen in prior studies. Consequently, answer correctness becomes a more accurate reflection of perception quality, which forms the theoretical foundation of the proposed AP-GRPO. Compared with vanilla GRPO, which assigns equal weight to all samples, AP-GRPO has several significant advantages: (a) It enhances visual understanding and feature extraction by encouraging attention to the correct ROIs within the image. (b) The model's grounding capabilities are directly optimized without relying on additional visual supervision, since AP-GRPO allows the reward signal to evaluate the grounding performance. (c) The proposed strategy also ensures interpretable reasoning.

Although AP-GRPO enables the model to learn precise visual localization, withholding full visual information inevitably causes a decrease in answer accuracy. Building upon the RL strategy, we further apply SFT to enhance the model's high-resolution reasoning capabilities. In the SFT phase, the original image is fully visible to the model. Following [4], we adopt a clean dataset separation: $\mathcal{D}_{\text{SFT}}$ for SFT and $\mathcal{D}_{\text{RL}}$ for RL. The loss function is represented as

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{T} \log P(y_t|x, y_{<t}; \theta), \tag{8}$$

where T is the text length and (x, y) represents the query and target response in dataset $\mathcal{D}_{\text{SFT}}$.

5 Experiments

Benchmarks. We evaluate the proposed method on several benchmarks targeting high-resolution visual understanding capabilities of LMMs. (a) The MME-RealWorld dataset [65] comprises challenging visual question-answering pairs, with an average resolution of $2,076 \times 1,434$. We use its training set for post-training, randomly sampling 10K examples for $\mathcal{D}_{\text{RL}}$ and assigning the remainder to $\mathcal{D}_{\text{SFT}}$. Its test set, also referred to as MME-RealWorld-Lite, contains

Table 2: Answer accuracy of state-of-the-art models on the in-distribution dataset MME-RealWorld-Lite [65]. **Bold** and underlined indicate the best and second best results respectively. The numbers in parentheses indicate the number of samples for each sub-task. Abbreviations: RS-Remote Sensing; MO-Monitoring; DT-Diagram and Table; AD-Autonomous Driving; OCR-Optical Character Recognition in the Wild.

Method	Parameters	Perception					Reasoning				Overall
		AD (350)	MO (319)	OCR (250)	RS (150)	DT (100)	AD (400)	DT (100)	OCR (100)	MO (150)	
Open-source General Models											
Qwen2.5-VL-7B [3]	7B	30.0	27.3	87.6	32.7	83.0	23.0	62.0	72.0	28.7	42.3
LLaVA-OneVision-7B [20]	7B	39.4	31.7	80.0	40.0	56.0	32.0	33.0	65.0	38.0	43.7
InternVL3-8B [68]	8B	36.9	34.5	83.6	49.3	75.0	37.0	44.0	70.0	40.0	47.9
Qwen2.5-VL-7B with Post-Training											
SFT	7B	43.7	43.6	85.6	55.3	78.0	41.0	60.0	70.0	50.7	54.0
GRPO [40]	7B	43.7	42.9	81.6	51.3	75.0	38.3	53.0	67.0	43.3	51.3
GRPO + SFT	7B	49.7	47.0	89.2	55.3	87.0	39.0	74.0	**77.0**	60.0	58.1
MGPO [16]	7B	44.0	46.7	86.4	54.0	78.0	39.3	69.0	74.0	52.7	55.1
MGPO + SFT	7B	55.4	49.8	83.6	**59.3**	82.0	47.8	71.0	71.0	**63.3**	60.5
Visual Grounded Reasoning Models											
Pixel-Reasoner-7B [42]	7B	30.9	38.9	89.6	52.0	86.0	32.5	72.0	71.0	46.0	49.7
DeepEyes-7B [67]	7B	33.4	43.3	**90.0**	52.7	**89.0**	35.0	69.0	76.0	44.0	53.2
HART-7B (Ours)	7B	**57.7**	**49.8**	89.6	58.7	86.0	**51.0**	**75.0**	72.0	58.7	**62.4**
Larger-scale Open-source Models											
Qwen2.5-VL-32B [3]	32B	40.7	29.5	87.2	40.7	83.0	29.5	60.0	74.0	27.3	45.6
InternVL3-38B [68]	38B	40.0	42.6	85.6	56.0	71.0	35.0	45.0	77.0	47.3	51.0
Qwen2.5-VL-72B [3]	72B	30.6	27.9	90.8	34.0	87.0	25.5	61.0	74.0	26.7	43.7
LLaVA-OneVision-72B [20]	72B	40.0	37.9	79.2	50.7	67.0	39.3	41.0	76.0	38.7	48.7
Private Models											
GPT-4o-mini [35]	—	24.2	26.5	62.5	6.7	44.2	26.8	39.1	47.0	25.8	36.4
Gemini-1.5-Pro [46]	—	26.6	31.1	67.6	14.0	39.9	19.2	33.2	52.7	28.3	38.2
GPT-4o [17]	—	22.4	33.9	77.7	28.9	46.7	26.4	44.8	61.4	36.5	45.2
Claude 3.5 Sonnet [1]	—	40.8	32.2	72.5	25.7	67.4	31.9	61.2	61.9	41.8	51.6

1,919 samples and is used for in-distribution evaluation. (b) TreeBench [48] serves as an out-of-distribution benchmark with an average image resolution of $2,152 \times 1,615$. Specifically, TreeBench contains a total of 833 manually annotated bounding boxes, providing a basis for evaluating grounding capabilities. Each sample in the datasets follows a multiple-choice format. The evaluation metric is the multiple-choice accuracy.

Baselines. We compare HART with several state-of-the-art baselines. (a) Private models including GPT-4o [17], Gemini series [9], etc. (b) Open-source general models including Qwen2.5-VL series [3], LLaVA-OneVision series [20], and InternVL3 series [68]. (c) Representative visual grounded reasoning models including Pixel-Reasoner [42] and DeepEyes [67]. (d) Post-training methods including GRPO [40] and the recently proposed MGPO [16] for high-resolution vision-centric tasks.

Implementation Details. We employ the Transformer Reinforcement Learning (TRL) framework [53] to enable distributed training and use Qwen2.5-VL-7B [3] as the base model. All reinforcement fine-tuning methods employ a binary reward function that evaluates answer correctness. For fair comparison, they are trained using the same training set of MME-RealWorld. The hyperparameter k in AP-GRPO is set to 0.6. Training details and cost are shown in Appendix C.

5.1 Main Results

We evaluate the answer accuracy of our proposed method HART on popular high-resolution visual benchmarks. Tab. 2 presents the accuracy comparison be-

Table 3: Answer accuracy of state-of-the-art models on the out-of-distribution dataset TreeBench [48]. **Bold** and underlined indicate the best and second best results respectively. The numbers in parentheses indicate the number of samples for each sub-task.

Method	Parameters	*Attributes (29)*	*OCR (68)*	*Material (13)*	*Obj. Retr. (16)*	*Phy. State (23)*	*Comparison (44)*	*Ordering (57)*	*Con. & Oc. (41)*	*Spa. Cont. (29)*	*Per. Trans. (85)*	**Overall**
		Perception					Reasoning					
Open-source General Models												
Qwen2.5-VL-7B [3]	7B	55.2	27.9	53.8	62.5	56.5	43.2	35.1	39.0	44.8	20.0	37.0
LLaVA-OneVision-7B [20]	7B	55.2	32.4	53.8	50.0	56.5	36.4	22.8	41.5	**72.4**	21.2	37.3
InternVL3-8B [68]	8B	51.7	33.7	**69.2**	56.3	56.5	43.2	24.6	39.0	72.4	**21.2**	38.8
Qwen2.5-VL-7B with Post-Training												
GRPO [40]	7B	44.8	51.5	46.2	50.0	56.5	47.7	28.1	41.5	44.8	14.1	38.0
GRPO + SFT	7B	48.3	47.1	61.5	50.0	52.2	43.2	28.1	43.9	55.2	15.3	38.5
MGPO [16]	7B	51.7	48.5	61.5	68.8	56.5	43.2	35.1	43.9	44.8	11.8	39.5
MGPO + SFT	7B	58.6	48.5	61.5	56.3	52.2	40.9	31.6	**51.2**	51.7	14.1	40.3
Visual Grounded Reasoning Models												
DeepEyes-7B [67]	7B	62.1	51.5	53.8	**68.8**	65.2	47.7	24.6	36.6	51.7	11.8	37.5
Pixel-Reasoner-7B [42]	7B	58.6	48.5	61.5	50.0	**65.2**	40.9	31.6	39.0	44.8	14.1	39.0
HART-7B (Ours)	7B	**62.1**	**55.9**	61.5	56.3	52.2	**50.0**	**35.1**	48.8	62.1	14.1	**43.7**
Larger-scale Open-source Models												
Qwen2.5-VL-32B [3]	32B	51.7	54.4	53.8	62.5	69.6	38.6	33.3	46.3	62.1	16.5	42.5
InternVL3-38B [68]	38B	51.7	51.5	61.5	68.8	52.2	38.6	33.3	56.1	65.5	12.9	42.0
Qwen2.5-VL-72B [3]	72B	65.5	48.5	69.2	56.3	56.5	38.6	33.3	51.2	72.4	11.8	42.2
LLaVA-OneVision-72B [20]	72B	62.1	36.8	53.8	62.3	65.2	47.7	28.1	53.7	65.5	12.9	40.5
Private Models												
Gemini-2.5-Flash [9]	—	48.3	75.0	53.9	68.8	69.6	43.2	19.3	56.1	72.4	15.3	45.9
GPT-4o [17]	—	51.7	69.1	61.5	43.8	65.2	43.2	38.6	48.8	72.4	18.8	46.9
Gemini-2.5-Pro [10]	—	51.7	83.8	61.5	75.0	56.5	54.6	36.8	65.9	86.2	20.0	54.1
o3 [36]	—	69.0	79.4	69.2	68.8	65.2	50.0	38.6	61.0	86.2	22.4	54.8

Table 4: Answer accuracy on other multimodal benchmarks.

Capability	Benchmark	Qwen2.5-VL-7B [3]	InternVL3-8B [68]	LLaVA-OneVision-7B [20]	**HART-7B (Ours)**
Multi-image VQA	BLINK [8]	55.2	55.5	48.2	**56.8**
	Mantis-Eval [19]	70.8	70.1	64.2	**72.8**
Low-resolution VQA	MathVista [30]	68.2	71.6	58.3	**71.8**
	MMStar [5]	59.3	53.5	56.7	**62.8**
High-resolution VQA	V* Bench [54]	78.5	72.3	70.7	**80.6**
	HR-Bench-4K [51]	70.1	70.8	64.3	**71.1**
	HR-Bench-8K [51]	61.0	62.0	59.8	**71.9**

tween HART and baselines on the in-distribution dataset. HART achieves an accuracy of 62.4% on MME-RealWorld-Lite, surpassing existing visual grounded reasoning models such as Pixel-Reasoner [42] and DeepEyes [67]. Notably, our HART outperforms the representative private and open-source models in most perception and reasoning tasks. Compared to the base model Qwen2.5-VL-7B [3], the proposed method achieves remarkable improvements on challenging tasks, *i.e.*, +26.0% on Remote Sensing, +27.7% on Perception-Autonomous Driving, and +30.0% on Reasoning-Monitoring tasks, indicating enhanced fine-grained visual understanding. We can observe that visual grounding methods are able to improve the model's performance on high-resolution visual tasks. Among them, HART achieves the highest average accuracy, demonstrating the effectiveness of our closed-loop framework. Furthermore, we apply HART to post-train the larger foundation model Qwen2.5-VL-32B and the recent model Qwen3-VL-8B. Detailed results are provided in Appendix D.

Table 5: Answer accuracy compared with Qwen2.5-VL-7B [3] and InternVL3-8B [68].

Method	HART (Ours)	MME-RealWorld-Lite [65]		TreeBench [48]	
		Perception	Reasoning	Perception	Reasoning
Qwen2.5-VL-7B [3]	✗	46.4	35.9	43.6	33.2
	✓	**64.9**	**58.5**	**57.0**	**35.9**
InternVL3-8B [68]	✗	51.1	42.9	46.3	34.4
	✓	**61.8**	**54.8**	**59.1**	**35.9**

Table 6: Grounding performance of AP-GRPO and baselines on the TreeBench [48] and Visual CoT [39] datasets. Bold numbers are the best results.

Method	TreeBench [48]		Visual-CoT [39]	
	Incorrect (↓)	Correct (↑)	Incorrect (↓)	Correct (↑)
Qwen2.5-VL-7B [3]	49.8%	50.2%	34.0%	66.0%
LLaVA-OneVision-7B [20]	61.7%	38.3%	33.0%	67.0%
InternVL3-8B [68]	84.9%	15.1%	71.1%	28.9%
Qwen2.5-VL-7B with Post-Training				
GRPO [40]	49.0%	51.0%	33.9%	66.1%
MGPO [16]	48.7%	51.3%	33.7%	66.3%
AP-GRPO (Ours)	**24.6%**	**75.4%**	**22.3%**	**77.7%**

Next, we evaluate the models on out-of-distribution datasets. As shown in Tab. 3, our HART also achieves open-source state-of-the-art on TreeBench with an accuracy of 43.7%. HART delivers significant improvements over the base model Qwen2.5-VL-7B [3], surpassing other post-training paradigms such as GRPO [40] and MGPO [16]. This demonstrates the effectiveness of the proposed post-training strategy in improving the joint understanding of visual and textual inputs. In Tab. 4, we compare HART with the base model Qwen2.5-VL-7B on a range of multimodal benchmarks covering both low- and high-resolution visual tasks. Detailed results are provided in Appendix D. The results highlight the strong adaptability of our method across visual scenes of different resolutions. Furthermore, we apply HART to post-train an additional base model InternVL3-8B [68]. As shown in Tab. 5, HART achieves better performance, indicating robust training behavior. In conclusion, our HART provides a flexible solution that adapts better to high-resolution real-world scenarios.

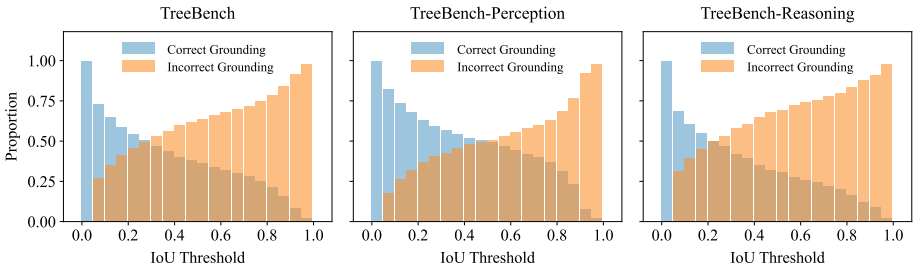

Fig. 3: Grounding performance of AP-GRPO on TreeBench [48].

Table 7: Ablations of each component of our HART.

Method	k	MME-RealWorld-Lite [65]		TreeBench [48]	
		Perception	Reasoning	Perception	Reasoning
Qwen2.5-VL-7B [3]	-	46.4	35.9	43.6	33.2
+ Post-Training					
GRPO [40]	-	52.1	49.1	52.3	32.0
MGPO [16]	-	58.0	50.5	53.7	31.2
HART-7B (Ours)					
· AP-GRPO	0.15	51.2	47.5	51.0	32.4
· AP-GRPO + SFT	0.15	**67.0**	56.9	56.4	31.6
· AP-GRPO	0.30	53.6	49.9	55.0	33.6
· AP-GRPO + SFT	0.30	64.1	56.0	**57.7**	34.8
· AP-GRPO	0.60	53.3	48.0	52.3	33.6
· SFT + AP-GRPO	0.60	51.7	48.1	55.0	32.4
· AP-GRPO + SFT	0.60	64.9	**58.5**	57.0	**35.9**

5.2 Grounding Results

In this section, we evaluate the grounding accuracy of the proposed method on the TreeBench [48] and Visual CoT [39] datasets, both of which provide ground-truth bounding box annotations for key image regions relevant to the questions. We utilize intersection over ground-truth as the metric to evaluate grounding accuracy. Tab. 6 presents the grounding performance with a coverage threshold of 0.3. AP-GRPO achieves superior grounding capabilities on both benchmarks, indicating its effectiveness in accurately localizing key regions. Compared to the base model Qwen2.5-VL-7B [3], the proposed method yields a +25.2% improvement on TreeBench [48] and a +11.7% improvement on Visual CoT [39]. While both AP-GRPO and MGPO adopt a visual grounded reasoning pipeline for post-training, the issue of reward misspecification limits MGPO's ability to deliver substantial improvements over the base model Qwen2.5-VL-7B. In contrast, HART can self-verify its localization results for policy updates, leading to more accurate localization. We report the grounding performance of AP-GRPO over a range of Intersection over Union (IoU) thresholds (0.0 to 0.95) in Fig. 3. These results suggest that AP-GRPO offers an effective optimization strategy for enhancing both grounding and reasoning capabilities. We evaluate grounding performance across different training stages and under more stringent intersection-over-ground-truth thresholds. Detailed results are shown in Appendix D.

5.3 Ablation Studies

In this experiment, we conduct ablation studies to investigate the influence of each component of HART. As demonstrated in Tab. 7, we compare the answer accuracy of post-trained Qwen2.5-VL-7B [3] with GRPO [40], MGPO [16], and the proposed HART on MME-RealWorld-Lite [65] and TreeBench [48], respectively. The results indicate that both RL (Stage 1) and SFT (Stage 2) are crucial for enhancing visual understanding, as HART yields considerably higher accuracy than either stage alone. In the table we also present the results of the sensitivity analysis on hyperparameter k. HART consistently improves performance under different values of k, demonstrating its robustness. Detailed results are provided in Appendix D.

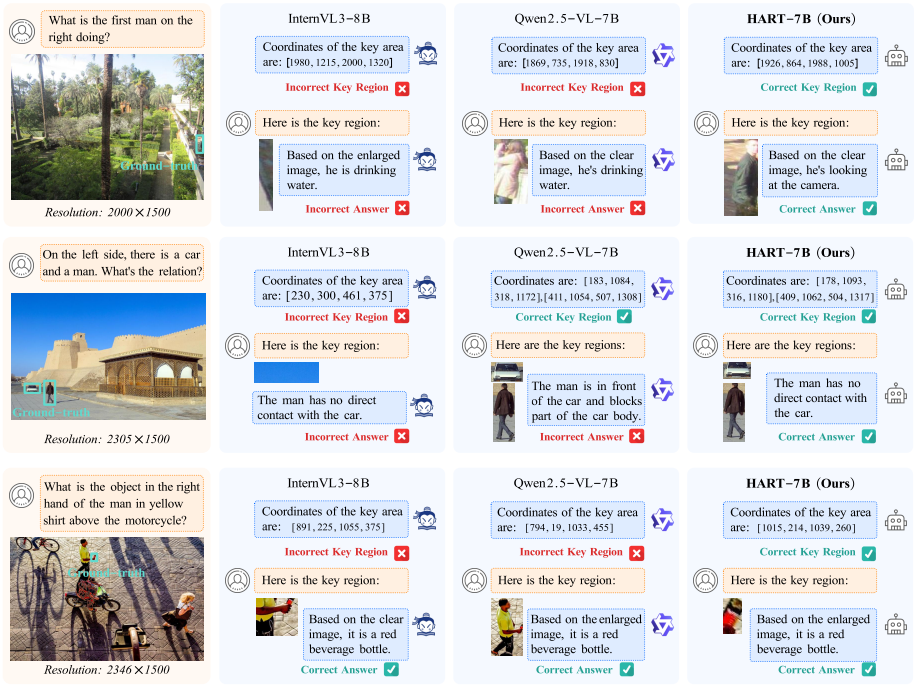

Fig. 4: Visualization of model outputs from InternVL3-8B [68], Qwen2.5-VL-7B [3], and our method HART-7B on TreeBench [48].

5.4 Visualization Results

We present a visualized comparison of our method against prior methods in Fig. 4. In this example, the task requires attending to the right-most man before executing the reasoning step. Compared with InternVL3-8B [68] and Qwen2.5-VL-7B [3], our method identifies the critical region more reliably and achieves a more accurate reasoning outcome. More failure cases and visualization results are shown in Appendices E and F.

6 Conclusion

This paper proposes HART, a closed-loop framework designed for high-resolution visual tasks. HART enables LMMs to identify and self-verify key regions of interest conditioned on visual and textual inputs. To guide its localization behavior, we propose AP-GRPO to prioritize optimizing samples with correct grounding behavior, without relying on external visual supervision. Empirical results demonstrate enhanced visual understanding across multiple high-resolution vision-centric tasks. In summary, HART provides a foundation for further exploration in the joint optimization of grounding and reasoning capabilities.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (62576160), Guangdong Basic and Applied Basic Research Foundation (2024A1515011340), the Fundamental Research Funds for the Central Universities (KG202508, KG202514), and 111 Center (B26023).

References

1. Anthropic: Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet (2024)
2. Arif, K.H.I., Yoon, J., Nikolopoulos, D.S., Vandierendonck, H., John, D., Ji, B.: Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1773–1781 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, L., Han, X., Shen, L., Bai, J., Wong, K.F.: Beyond two-stage training: Cooperative sft and rl for llm reasoning. arXiv preprint arXiv:2509.06948 (2025)
5. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
6. Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., et al.: Explainable and interpretable multimodal large language models: A comprehensive survey. arXiv preprint arXiv:2412.02104 (2024)
7. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Zhang, S., Duan, H., Zhang, W., Li, Y., et al.: Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. Advances in Neural Information Processing Systems **37**, 42566–42592 (2024)
8. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
9. Google DeepMind: Gemini-2.5-flash. https://deepmind.google/models/gemini/flash/ (2025)
10. Google DeepMind: Gemini-2.5-pro. https://deepmind.google/models/gemini/pro/ (2025)
11. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
12. Guo, Z., Hong, M., Jin, T.: Observe-r1: Unlocking reasoning abilities of mllms with dynamic progressive reinforcement learning. arXiv preprint arXiv:2505.12432 (2025)
13. Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Huang, G.: Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In: European Conference on Computer Vision. pp. 390–406. Springer (2024)
14. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. CoRR **abs/2403.12895** (2024)

15. Huang, R., Ding, X., Wang, C., Han, J., Liu, Y., Zhao, H., Xu, H., Hou, L., Zhang, W., Liang, X.: Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29814–29824 (2025)
16. Huang, X., Dong, Y., Tian, W., Li, B., Feng, R., Liu, Z.: High-resolution visual reasoning via multi-turn grounding-based reinforcement learning. arXiv preprint arXiv:2507.05920 (2025)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
19. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483 (2024)
20. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
21. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multimodal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024)
22. Li, Z., Yu, W., Huang, C., Liu, R., Liang, Z., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H., et al.: Self-rewarding vision-language model via reasoning decomposition. arXiv preprint arXiv:2508.19652 (2025)
23. Liu, C., Yin, K., Cao, H., Jiang, X., Li, X., Liu, Y., Jiang, D., Sun, X., Xu, L.: Hrvda: High-resolution visual document assistant. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15534–15545 (2024)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
25. Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al.: Llava-plus: Learning to use tools for creating multimodal agents. In: European conference on computer vision. pp. 126–142. Springer (2024)
26. Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: Textmonkey: An ocr-free large multimodal model for understanding document. CoRR **abs/2403.04473** (2024)
27. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
28. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In: The Thirteenth International Conference on Learning Representations (2025)
29. Liu, Z., Dong, Y., Rao, Y., Zhou, J., Lu, J.: Chain-of-spot: Interactive reasoning improves large vision-language models. CoRR **abs/2403.12966** (2024)
30. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: International Conference on Learning Representations. vol. 2024, pp. 23439–23554 (2024)

31. Luan, B., Feng, H., Chen, H., Wang, Y., Zhou, W., Li, H.: Textcot: Zoom in for enhanced multimodal text-rich image understanding. CoRR **abs/2404.09797** (2024)
32. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
33. Miri Rekavandi, A., Xu, L., Boussaid, F., Seghouane, A.K., Hoefs, S., Bennamoun, M.: A guide to image- and video-based small object detection using deep learning: Case study of maritime surveillance. IEEE Transactions on Intelligent Transportation Systems **26**(3), 2851–2879 (2025)
34. Ni, M., Yang, Z., Li, L., Lin, C.C., Lin, K., Zuo, W., Wang, L.: Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. arXiv preprint arXiv:2505.19702 (2025)
35. OpenAI: Gpt-4o mini: advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/` (2024)
36. OpenAI: Openai-o3. `https://openai.com/index/introducing-o3-and-o4-mini/` (2025)
37. Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)
38. Ptito, M., Bleau, M., Bouskila, J.: The retina: a window into the brain (2021)
39. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: 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)
40. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
41. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
42. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025)
43. Sun, G., Jin, M., Wang, Z., Wang, C.L., Ma, S., Wang, Q., Geng, T., Wu, Y.N., Zhang, Y., Liu, D.: Visual agents as fast and slow thinkers. arXiv preprint arXiv:2408.08862 (2024)
44. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
45. Tan, H., Ji, Y., Hao, X., Lin, M., Wang, P., Wang, Z., Zhang, S.: Reason-rft: Reinforcement fine-tuning for visual reasoning. arXiv preprint arXiv:2503.20752 (2025)
46. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
47. Thawakar, O., Dissanayake, D., More, K., Thawkar, R., Heakl, A., Ahsan, N., Li, Y., Zumri, M., Lahoud, J., Anwer, R.M., et al.: Llamav-o1: Rethinking step-by-step visual reasoning in llms. CoRR (2025)
48. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., et al.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. arXiv preprint arXiv:2507.07999 (2025)

49. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)
50. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
51. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: 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. vol. 39, pp. 7907–7915 (2025)
52. Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H.: Multimodal chain-of-thought reasoning: A comprehensive survey. CoRR **abs/2503.12605** (March 2025)
53. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl` (2020)
54. Wu, P., Xie, S.: V^*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
55. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
56. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., et al.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
57. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Minicpm-v: A gpt-4v level mllm on your phone. CoRR **abs/2408.01800** (2024)
58. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024)
59. Yu, Z., Huang, H., Chen, W., Su, Y., Liu, Y., Wang, X.: Yolo-facev2: A scale and occlusion aware face detector. Pattern Recognition **155**, 110714 (2024)
60. Zhan, Y., Zheng, S., Zhu, Y., Zhao, H., Yang, F., Tang, M., Wang, J.: Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22947–22957 (2025)
61. Zhan, Y., Zhu, Y., Zheng, S., Zhao, H., Yang, F., Tang, M., Wang, J.: Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. CoRR (2025)
62. Zhang, J., Huang, J., Yao, H., Liu, S., Zhang, X., Lu, S., Tao, D.: R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. CoRR **abs/2503.12937** (March 2025)
63. Zhang, R., Shao, R., Chen, G., Zhang, M., Zhou, K., Guan, W., Nie, L.: Falcon: Resolving visual redundancy and fragmentation in high-resolution multimodal large language models via visual registers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23530–23540 (2025)
64. Zhang, Y.F., Wen, Q., Fu, C., Wang, X., Zhang, Z., Wang, L., Jin, R.: Beyond llava-hd: Diving into high-resolution large multimodal models. CoRR **abs/2406.08487** (2024)

65. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257 (2024)
66. Zhao, K., Zhu, B., Sun, Q., Zhang, H.: Unsupervised visual chain-of-thought reasoning via preference optimization. CoRR **abs/2504.18397** (April 2025)
67. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
68. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)