Learning Active Perception for Pixel-Space Reasoning via Visual-Intent Stratified GRPO

Mingkang Zhu[1] Xi Chen[3] Senqiao Yang[1]
Bei Yu[1] Hengshuang Zhao[3] Jiaya Jia[2]

[1]CUHK [2]HKUST [3]HKU

Abstract. Pixel-space reasoning enables Vision-Language Models to perform **active perception**: instead of answering from a single global view, the model can zoom into high-resolution regions to gather fine-grained evidence. This mirrors human visual problem solving, which involves broad exploration (finding relevant cues) and local verification (confirming details). However, most pixel-space reasoning systems are trained with Group Relative Policy Optimization (GRPO), whose group-wise advantage normalization directly compares sampled rollouts that follow different perception strategies. Rollouts with different zoom depths and different visual intents (explore vs. verify) are normalized together, making exploratory multi-zoom behavior appear disadvantageous: the multi-zoom strategies newly discovered during training typically yield lower mean rewards than already-mastered single-zoom or no-zoom strategies. This leads to **visual laziness**: the policy does not explore active multi-zoom perception strategies and collapses into a zoom-averse state. To address this, we propose **Visual-Intent Stratified GRPO (VIS-GRPO)**, a drop-in replacement for GRPO that restores fair learning signals for active perception. VIS-GRPO computes advantages only among strategically comparable rollouts by stratifying trajectories along (i) **zoom depth** and (ii) **visual intent**, separating broad search over diverse regions from local verification over overlapping regions. This alignment prevents exploratory trajectories from being overshadowed and encourages the learning of active perception. Extensive experiments demonstrate that VIS-GRPO enables effective multi-zoom active perception strategies and consistently improves performance across challenging visual understanding benchmarks like HR-Bench and MME-RealWorld.

Keywords: Pixel-Space Reasoning · Vision-Language Models · Reinforcement Learning

1 Introduction

Vision–Language Models (VLMs) have recently achieved strong performance across a broad range of visual understanding and reasoning tasks, driven by advances in post-training with reinforcement learning (RL) [4,8,10,12,18,21,23,31]. Most existing RL-based VLM post-training frameworks formulate multimodal

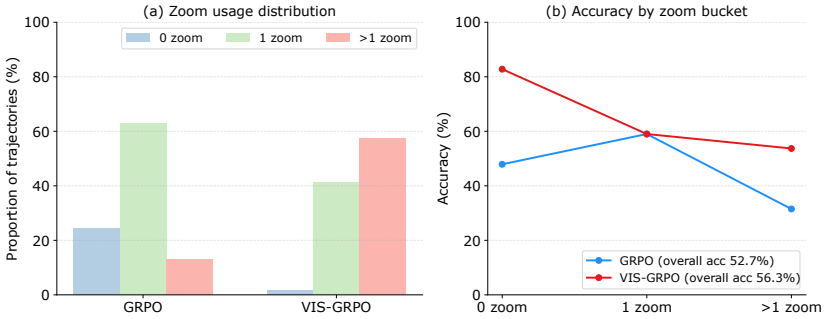


Fig. 1: Observation on MME-RealWorld. (a) Distribution of zoom usage for GRPO vs. VIS-GRPO; bar heights indicate the fraction of test questions whose sampled response uses 0 / 1 / >1 zoom. (b) Accuracy of responses within each zoom bucket. GRPO exhibits a strong preference for low-zoom behaviors and performs poorly when performing multi-zoom operations, reflecting **visual laziness**. In contrast, VIS-GRPO substantially increases multi-zoom usage and improves overall accuracy.

reasoning as text generation conditioned on a fixed visual representation and optimize the policy primarily within the textual domain. While effective in many settings, this "single-look" paradigm can be fundamentally limited when handling high-resolution and visually dense inputs, where correct answers depend on fine-grained, localized evidence such as small text, subtle symbols, or tiny objects.

Recently, pixel-space reasoning has emerged as a promising paradigm [14, 24, 40, 41]. Instead of relying on a single global encoding of the image, pixel-space reasoning interleaves language generation with explicit zoom-in actions on the original image. Each zoom-in extracts a high-resolution crop from a selected region, appends the new observation to the context, and enables subsequent reasoning conditioned on the newly acquired evidence. This process transforms visual reasoning into **active perception**, where the model must decide where to look and what to verify in order to answer correctly.

Motivation. However, as shown in Fig. 1 and Fig. 4, we observe that under standard GRPO training, as adopted in previous work [24, 41], pixel-space reasoning models often exhibit **visual laziness**. Concretely, the learned policy does not proactively utilize the zoom-in operator and, in some cases, collapses to rarely invoking it, even when fine-grained evidence is required to solve the task. Furthermore, when a GRPO-trained model produces multi-zoom trajectories, **the accuracy is noticeably lower** than trajectories using zero or one zoom operation, indicating that GRPO not only discourages multi-step evidence gathering but also fails to reliably learn effective multi-zoom strategies. These observations suggest that standard GRPO training does not fully realize the goal of **active perception** in pixel-space reasoning—learning to iteratively search for and verify localized evidence in high-resolution, visually complex images. This raises a natural question: **why does GRPO induce visual laziness in pixel-space**

reasoning, and how can we train models to gain strong multi-step active perception ability?

Visual Laziness. Compared to the complicated multi-zoom strategies newly discovered during RL training, the model is more competent at simple no-zoom or single-zoom strategies mastered during pretraining; group-relative normalization then assigns systematically negative advantages to exploratory multi-zoom rollouts, reducing their probabilities of being sampled before they can be properly learned. This creates a self-reinforcing loop that induces **visual laziness**: the policy increasingly avoids zoom-in operations and fails to learn effective active perception.

Ineffective Multi-zoom Strategies. We argue that this phenomenon reflects a structural mismatch between GRPO's group-relative advantage normalization and the strategy space of pixel-space reasoning trajectories. In pixel-space reasoning, rollouts are heterogeneous along two strategy-defining axes: (i) **zoom depth**, i.e., the number of zoom-in operations, which determines how much high-resolution evidence is acquired; and (ii) **visual intent**, i.e., whether zoom-ins are used for broad search across diverse regions or for local verification of subtle details within previously zoomed areas. Standard GRPO normalizes advantages by directly comparing such heterogeneous strategies within the same rollout group. As a result, the model tends to over-optimize simpler no-zoom or single-zoom behaviors with higher immediate reward, while under-optimizing multi-zoom strategies whose rewards are initially noisy due to imperfect zoom region selection, thereby causing a drop in multi-zoom performance.

To address these challenges, we propose **Visual-Intent Stratified GRPO (VIS-GRPO)**, a drop-in replacement for GRPO tailored to pixel-space reasoning. The core principle is that advantages should be computed only among rollouts that follow comparable perception strategies. VIS-GRPO implements this principle by stratifying sampled trajectories along (i) zoom depth and (ii) visual intent. Within the multi-zoom regime, we define visual intent using a lightweight spatial coherence statistic based on the mean pairwise IoU among zoom regions: high coherence corresponds to `verify` (local verification), and low coherence corresponds to `explore` (broad search). GRPO advantages are then computed and normalized **within each depth–intent stratum**, preventing exploratory active-perception behaviors from being unfairly overshadowed by low-zoom behaviors. VIS-GRPO requires only trajectory-level statistics already available during training and can be integrated into existing pixel-space reasoning pipelines without architectural changes.

Empirically, VIS-GRPO leads to effective multi-zoom strategies and consistently improves pixel-space reasoning across diverse challenging benchmarks. Notably, VIS-GRPO improves the base model's performance by 8.7% on HR-Bench 8K and 7.6% on MME-RealWorld and surpasses baseline pixel-reasoning systems. Our main contributions are:

- We identify **visual laziness** as a key failure mode when training pixel-space reasoning with GRPO, and provide a principled analysis showing that group-

relative advantage normalization is misaligned with heterogeneous zoom-based perception strategies.
- We propose **VIS-GRPO**, a drop-in RL algorithm that stratifies rollouts by **zoom depth** and **visual intent** and computes GRPO advantage within each depth–intent stratum, ensuring fair comparisons among strategically comparable trajectories.
- Extensive experiments show that VIS-GRPO learns effective active perception strategies and achieves consistent performance improvement on challenging visual reasoning benchmarks.

2 Related Work

Reinforcement Learning for Vision-Language Models. Reinforcement learning (RL) [5, 13, 22, 25, 32, 43, 44] has been widely adopted for post-training Large Language Models (LLMs) [1, 7, 20, 21, 36, 42]. Building on these advances, recent efforts have adapted these methods to the vision-language domain, particularly to enhance the reasoning abilities of Vision-Language Models (VLMs) [4, 8, 10, 12, 18, 31]. For example, MM-Eureka [18] expands training data coverage across multiple domains. VLAA-Thinking [4] carefully analyzes the SFT and RL stages for visual reasoning and curates a dataset for related tasks. ThinkLite [31] improves data efficiency via sample selection. Despite these advancements, a common limitation is that the reasoning process in these models is confined to the textual domain, leaving the visual input itself untouched. This lack of direct interaction motivates pixel-space reasoning, which aims to actively acquire additional visual evidence from the input images.

Pixel-Space Reasoning for Vision-Language Models. Pixel-space reasoning trains VLMs to interleave textual chain-of-thought reasoning with iterative zoom-ins on image regions, enabling an **active perception** loop for evidence acquisition [14, 24, 40, 41]. This paradigm has shown strong potential in challenging high-resolution and visually complex tasks where fine-grained local evidence is critical. However, existing pixel-space reasoning systems are almost universally trained with GRPO, under which zoom usage often decreases during training, and the policy may fail to learn reliable multi-zoom perception behaviors. Although some prior works introduce additional rewards to encourage zoom-based trajectories [24], these heuristics do not directly address the underlying algorithmic cause of this degeneration. Our work analyzes and targets the underlying mismatch between GRPO's group-relative advantage normalization and heterogeneous zoom strategies, and proposes a training algorithm that is better aligned with the strategy space of zoom-based active perception.

3 Method

In this section, we present **Visual-Intent Stratified GRPO (VIS-GRPO)**, a reinforcement learning algorithm tailored to pixel-space reasoning. We first formulate the problem of pixel-space reasoning and review standard GRPO in

Sec. 3.1. In Sec. 3.2, we then analyze why GRPO induces visual laziness. Finally, we introduce VIS-GRPO as a solution in Sec. 3.3.

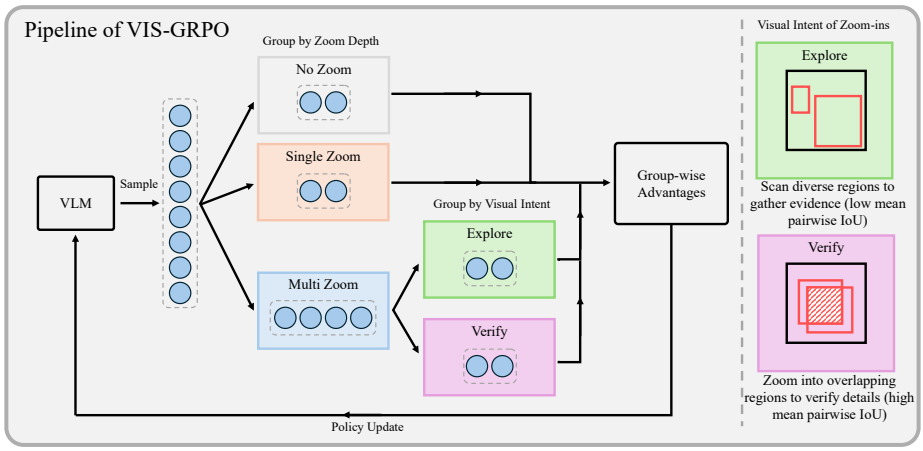


Fig. 2: Overall pipeline of VIS-GRPO. VIS-GRPO first groups rollouts by **zoom depth** (the number of zoom-in operations), avoiding direct comparison between trajectories that acquire different amounts of high-resolution evidence. Within the multi-zoom regime, VIS-GRPO further stratifies by **visual intent** using a spatial coherence score based on the mean pairwise IoU of zoom regions: high coherence indicates `verify` (local verification in nearby locations), while low coherence indicates `explore` (broad search over diverse regions). Group-relative advantages are computed and normalized within each depth-intent stratum.

3.1 Preliminaries

Problem Formulation. In pixel-space reasoning, VLMs interleave language token generation with pixel-space zoom-in operations that extract fine-grained details from the original image [14, 24, 41]. Let the vision-language input be $\mathbf{x} = [V, L]$, where V is the visual input and L is the textual prompt. A VLM policy π_θ generates a trajectory $\tau \sim \pi_\theta(\cdot \mid \mathbf{x})$ that alternates between textual reasoning and optional zoom-in actions. Each zoom-in action is parameterized by a normalized bounding box $b_t \in [0, 1]^4$ and invokes an image crop operator f to extract a high-resolution observation $\mathbf{o}_t = f(V, b_t)$, which is appended to the context for subsequent reasoning. The overall RL objective is:

$$\max_\theta \mathcal{J}(\theta) \;=\; \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\, \tau\sim\pi_\theta(\cdot|\mathbf{x})}[R(\tau)]. \tag{1}$$

Group Relative Policy Optimization (GRPO). GRPO [23] is a widely used RL algorithm for training VLMs and pixel-space reasoning systems [24, 41]. For

a given input $\mathbf{x}$, GRPO samples a group of k trajectories $\{\tau_1, \ldots, \tau_k\}$ from the behavior policy $\pi_{\theta_{\text{old}}}$:

$$\tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{x}).$$

Let $R_i := R(\tau_i)$ denote the trajectory-level reward. GRPO computes a group-relative normalized advantage by standardizing rewards within the sampled group:

$$\tilde{A}_i = \frac{R_i - \bar{R}}{\sigma_R + \epsilon}, \qquad \bar{R} = \frac{1}{k}\sum_{j=1}^{k} R_j, \qquad \sigma_R^2 = \frac{1}{k}\sum_{j=1}^{k}(R_j - \bar{R})^2, \tag{2}$$

where $\epsilon > 0$ ensures numerical stability.

The policy is updated by maximizing a clipped surrogate objective with KL regularization against a fixed reference model π_{ref}:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},\{\tau_i\}_{i=1}^{k}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{x})}\Bigg[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\Big\{\min\Big(\rho_{i,t}(\theta)\tilde{A}_i,\\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\tilde{A}_i\Big) - \beta\,\mathbb{D}_{\text{KL}}\big(\pi_\theta(\cdot \mid \mathbf{x}, \tau_{i,<t}) \,\|\, \pi_{\text{ref}}(\cdot \mid \mathbf{x}, \tau_{i,<t})\big)\Big\}\Bigg],\end{aligned} \tag{3}$$

where $\rho_{i,t}(\theta) = \frac{\pi_\theta(\tau_{i,t}|\mathbf{x},\tau_{i,<t})}{\pi_{\theta_{\text{old}}}(\tau_{i,t}|\mathbf{x},\tau_{i,<t})}$.

3.2 Analysis of GRPO's Shortcomings in Pixel-Space Reasoning

Group Relative Policy Optimization (GRPO) [23] has shown strong performance in text-only reasoning, largely because group-wise advantage normalization can provide a stable learning signal when rollouts within a group are strategically comparable [7,23,37]. However, pixel-space reasoning violates this comparability assumption: rollouts for the same prompt can follow qualitatively different **perception strategies**. As a result, GRPO's group-relative comparison can distort credit assignment across strategies and systematically suppress the learning of effective multi-step zoom behaviors, leading to **visual laziness** failure mode highlighted in Fig. 1 and Fig. 4.

Pixel-Space Reasoning Induces Structured Heterogeneity. Pixel-space reasoning turns visual understanding into **active perception**: the model chooses **where to look** and **what to verify** via zoom-in actions. As a result, trajectories exhibit two structured axes of heterogeneity: (i) **zoom depth**—the number of zoom-in operations $c(\tau)$, which determines the amount of high-resolution evidence a trajectory collects; and (ii) **visual intent**—how zoom-ins are used. Even with the same zoom depth, multi-zoom trajectories can serve different intents: `explore` rollouts distribute zooms across diverse regions to search for relevant cues, whereas `verify` rollouts repeatedly focus on overlapping regions to confirm subtle details.

These axes induce distinct perception strategies. Comparing these diverse strategies in advantage estimation makes it hard for the model to distinguish

whether the received reward should be attributed to the choice of certain perception strategies or the execution quality of such strategies.

GRPO Induces Visual Laziness. Under GRPO, the normalized advantage for a rollout is computed relative to other rollouts sampled for the same prompt, as shown in Eq. (2). When a group mixes fundamentally different perception strategies, this relative comparison can become systematically biased during learning. In particular, no-zoom or single-zoom strategies are more familiar to the base model. They are often easier to execute and can yield higher immediate rewards. In contrast, multi-zoom strategies are often newly discovered during RL training and require multiple accurate, consistent region selections and longer-horizon evidence integration. Early on, the model is not yet proficient at these operations, so multi-zoom rollouts are more error-prone and receive lower rewards even when multi-step zooming is essential for solving high-resolution complex visual understanding tasks.

As a consequence, GRPO's group-relative normalization assigns systematically negative advantages to exploratory multi-zoom rollouts. These negative learning signals suppress the updates needed to improve region selection and multi-turn evidence gathering, causing multi-zoom behaviors to be sampled less frequently and to remain under-trained. This creates a self-reinforcing loop in which the policy increasingly prefers low-zoom behaviors and gradually collapses toward zoom-averse solutions. We refer to this failure mode as **visual laziness**: the policy avoids active evidence acquisition and answers from coarse visual information even when fine-grained details are required. This behavior is consistent with our observations in Fig. 1 and Fig. 4, where zoom usage decreases during GRPO training and multi-zoom strategies remain both rare and inaccurate.

Controlling Zoom Depth Alone is Insufficient. A reasonable first step is to avoid comparing rollouts with different zoom depths, so that strategies with different evidence budgets are not normalized against each other. However, pixel-space reasoning also exhibits a visual-intent heterogeneity in the multi-zoom regime: `explore` and `verify` correspond to qualitatively different visual intents with distinct spatial patterns. As illustrated in Fig. 5, if these intents are still mixed during advantage normalization, learning can become unbalanced, weakening either the broad search and the local verification behaviors required for active perception.

The above analysis suggests that the core issue is structural: GRPO performs group-relative normalization across rollouts that are not strategically comparable in pixel-space reasoning. These observations motivate our approach: we align advantage estimation with the intrinsic structure of pixel-space reasoning by stratifying rollouts by both zoom depth and visual intent before computing group-relative advantages.

3.3 Visual-Intent Stratified Grouping

Our solution is based on the principle that **advantages should be computed only among rollouts that are strategically comparable**. We therefore

propose **Visual-Intent Stratified Grouping (VISG)**, which partitions trajectories into homogeneous strata. Applying group-wise advantage normalization after VISG yields our full algorithm, **Visual-Intent Stratified GRPO (VIS-GRPO)**, as a drop-in replacement for standard GRPO.

Stratification Axes. VISG stratifies trajectories along two axes that capture the dominant sources of heterogeneity in pixel-space reasoning:

1. **Zoom depth (evidence budget).** We measure zoom depth by the number of zoom-in operations $c(\tau)$ in trajectory τ. Different $c(\tau)$ corresponds to different amounts of high-resolution evidence, making direct comparison unreliable.
2. **Visual intent (explore vs. verify).** Depth alone is insufficient: two trajectories with the same $c(\tau)$ can pursue distinct intents: `explore` performs broad search over diverse regions, while `verify` repeatedly focuses on overlapping regions to confirm subtle details. Mixing these intents reintroduces heterogeneity even within the same depth level.

Step 1: Stratify by Zoom Depth. Given a prompt $\mathbf{x}$ and k sampled rollouts $\{\tau_i\}_{i=1}^{k}$, we first partition them by zoom depth $c(\tau_i)$. This prevents low-zoom strategies (which are often easier early in training) from dominating the advantage normalization baseline for multi-zoom rollouts.

Step 2: Stratify Multi-zoom Rollouts by Visual Intent. Within each depth group where $c(\tau) \geq 2$, we further partition rollouts into `explore` vs. `verify` using a lightweight spatial coherence statistic. Specifically, let $\{b_1, \ldots, b_{c(\tau)}\}$ be the sequence of zoom bounding boxes in trajectory τ. We compute the mean pairwise IoU:

$$\overline{\mathrm{IoU}}(\tau) \;=\; \frac{2}{c(\tau)\big(c(\tau)-1\big)} \sum_{1 \leq i < j \leq c(\tau)} \mathrm{IoU}(b_i, b_j), \tag{4}$$

which measures intra-trajectory overlap. A high $\overline{\mathrm{IoU}}(\tau)$ indicates repeated focus on similar regions, consistent with `verify`; a low $\overline{\mathrm{IoU}}(\tau)$ indicates coverage of disparate regions, consistent with `explore`. With a threshold β, we assign an intent label

$$m(\tau) = \begin{cases} \texttt{verify}, & \overline{\mathrm{IoU}}(\tau) \geq \beta, \\ \texttt{explore}, & \overline{\mathrm{IoU}}(\tau) < \beta, \\ \texttt{none}, & c(\tau) < 2. \end{cases}$$

Finally, each trajectory is assigned to a discrete stratum $s(\tau) = (c(\tau), m(\tau))$.

Stratum-wise Advantage Normalization. VIS-GRPO replaces the group-wise statistics in Eq. (2) with stratum-wise statistics. For a fixed prompt $\mathbf{x}$ and a stratum s, define the index set

$$\mathcal{I}_{\mathbf{x},s} = \{\, i \mid s(\tau_i) = s \,\}.$$

We compute the stratum mean and standard deviation of returns:

$$\bar{R}_{\mathbf{x},s} = \frac{1}{|\mathcal{I}_{\mathbf{x},s}|} \sum_{i \in \mathcal{I}_{\mathbf{x},s}} R_i, \qquad \sigma^2_{\mathbf{x},s} = \frac{1}{|\mathcal{I}_{\mathbf{x},s}|} \sum_{i \in \mathcal{I}_{\mathbf{x},s}} \big(R_i - \bar{R}_{\mathbf{x},s}\big)^2. \tag{5}$$

The normalized advantage for rollout τ_i is then

$$\tilde{A}_i^{\mathrm{VIS}} = \frac{R_i - \bar{R}_{\mathbf{x},s(\tau_i)}}{\sigma_{\mathbf{x},s(\tau_i)} + \epsilon}. \tag{6}$$

Compared to standard GRPO (Eq. (2)), the only change is that the baseline and standard deviation are computed using only the rollouts in the same depth–intent stratum. This enforces fair comparisons among rollouts that share both evidence budget (zoom depth) and perception purpose (visual intent), stabilizing the learning of multi-zoom strategies and preventing exploratory active-perception behaviors from being overshadowed by simple strategies. The full pipeline of VIS-GRPO is illustrated in Fig. 2.

Beyond the empirical motivation in Sec. 3.2, we provide a formal rationale and derived analysis in the supplementary material, which explains why visual-intent stratification is necessary for pixel-space reasoning and why controlling zoom depth alone can remain insufficient.

4 Experiments

This section is organized as follows. We first describe the experiment setup in Sec. 4.1. Next, in Sec. 4.2, we compare our VIS-GRPO against a diverse set of pixel-space reasoning systems and RL algorithms. Finally, we present qualitative case studies and conduct an in-depth analysis of VIS-GRPO.

4.1 Experiment Setup

Models and Training. For RL training, we use Pixel Reasoner's training set, which consists of 15,000 samples drawn from the Pixel Reasoner's SFT data, InfographicVQA [17], LLaVA-CoT [35], and STARQA [33]. We adopt Qwen3-VL-8B as the base model [2]. Exact Match (EM) is used as the reward for RL training. Additional experiment details are in the supplemental materials.

Evaluation Benchmarks. We evaluate our method and baselines on four representative visual understanding benchmarks: V* [34], HR-Bench 4K [29], HR-Bench 8K, and MME-RealWorld [38]. These benchmarks assess VLMs on their ability to understand high-resolution, visually complex images and identify fine-grained visual cues. We report accuracy for all benchmarks.

Baselines. We first compare the model trained with our VIS-GRPO with a broad set of baselines covering (i) general-purpose VLMs and (ii) pixel-space reasoning systems. General-purpose VLMs include GPT-4o [19], Gemini-2.0-Flash [27], Gemini-2.5-Pro [6,26], LLaVA-OneVision [15], DeepSeek-VL-7B [16], InternLM-XComposer2-4KHD (IXC2-4KHD) [9], Qwen2.5-VL-7B [3], Qwen2.5-VL-72B, Video-R1 [10], LongLLaVA [30], Gemma3-27B-IT [28], and Qwen3-VL-8B [2]. Pixel-space reasoning systems include Instruction-Guided Masking (IVM-Enhance) [39], Visual-Program Distillation (PaLI-X-VPD) [11], SEAL [34], DeepEyes [41], and Pixel Reasoner [24].

Table 1: Performance comparison of existing VLMs, pixel-space reasoning systems, and ours on visual reasoning benchmarks. Accuracy is reported for all benchmarks. † denotes methods using GPT-4V.

Model	Size	V* Bench	HR-Bench 4K	HR-Bench 8K	MME-RealWorld	Avg.
General-purpose VLMs						
GPT-4o	—	62.8	59.0	55.5	46.4	55.9
Gemini-2.0-Flash	—	73.2	—	—	—	—
Gemini-2.5-Pro	—	79.2	—	—	—	—
LLaVA-OneVision	7B	75.4	63.0	59.8	48.5	61.7
DeepSeek-VL-7B	7B	—	35.5	33.4	—	—
IXC2-4KHD	7B	—	57.8	51.3	—	—
Video-R1	7B	51.2	—	—	—	—
LongLLava	13B	68.5	—	—	—	—
Gemma3-27B-IT	27B	62.3	—	—	—	—
Qwen2.5-VL-72B	72B	69.1	67.6	68.0	—	—
Qwen2.5-VL-7B	7B	73.3	67.3	64.1	44.6	62.3
Qwen3-VL-8B	8B	86.4	77.8	72.9	48.7	71.5
Pixel-space Reasoning Systems						
IVM-Enhance†	—	81.2	—	—	—	—
SEAL	7B	74.8	—	—	—	—
PaLI-X-VPD	55B	76.6	—	—	—	—
Pixel Reasoner	7B	84.3	72.6	66.1	—	—
DeepEyes	7B	82.7	74.3	68.8	51.9	69.4
VIS-GRPO	8B	**88.5**	**81.9**	**81.6**	**56.3**	**77.1**

RL Algorithm Baselines. To isolate the effect of the RL training algorithm, we additionally compare VIS-GRPO with representative RL algorithms in pixel-space reasoning under a controlled setup: all algorithms start from the same base model (Qwen3-VL-8B) and use the same training data. We consider: (i) **GRPO** [23], the standard GRPO algorithm widely used in prior pixel-space reasoning systems; (ii) **Pixel Reasoner** [24], which additionally introduces specific reward heuristics to encourage zoom-based trajectories; and (iii) **Depth-Stratified GRPO**, a one-axis ablation that normalizes advantages only among rollouts with the same zoom depth. This depth-only stratification is related to stratified grouping based on the number of search queries used for LLM search agents [45]. Depth-Stratified GRPO removes the zoom-depth mismatch while leaving the visual-intent heterogeneity unmodeled. In contrast, VIS-GRPO explicitly addresses this missing dimension via visual-intent stratification.

4.2 Main Results

Comparison with VLMs and Pixel-Space Reasoning Systems. As shown in Tab. 1, applying VIS-GRPO to the Qwen-3-VL-8B model substantially improves its visual understanding capability. Our resulting model achieves an average score of 77.1, representing a nontrivial 5.6-point gain over the base model. This improvement is particularly pronounced on high-resolution benchmarks,

Table 2: Performance comparison of the Qwen3-VL-8B model trained with representative RL algorithms and our VIS-GRPO, evaluated on visual reasoning benchmarks. We report accuracy for all benchmarks.

Model	V* Bench	HR-Bench 4K	HR-Bench 8K	MME-RealWorld	Avg.
Qwen3-VL-8B	86.4	77.8	72.9	48.7	71.5
GRPO	87.4	79.0	77.6	52.7	74.2
Pixel Reasoner	83.8	80.6	79.0	54.2	74.4
Depth-Stratified GRPO	84.8	79.1	79.4	55.7	74.8
VIS-GRPO	**88.5**	**81.9**	**81.6**	**56.3**	**77.1**

which require fine-grained evidence acquisition via pixel-space reasoning. Specifically, VIS-GRPO yields a remarkable 8.7-point gain on HR-Bench 8K and a 7.6-point gain on MME-RealWorld. This strongly demonstrates the effectiveness of our approach in learning complex multi-zoom pixel-space reasoning strategies critical for high-resolution images. Overall, our model achieves the highest average score among all models in Tab. 1. These results validate that VIS-GRPO successfully learns effective pixel-space reasoning strategies that translate into stronger visual understanding.

Comparison with RL Training Algorithms. To isolate the effect of the RL algorithm, we compare VIS-GRPO with representative RL algorithms under the controlled setup described in Sec. 4.1: all methods start from the same base model and use the same training data. As reported in Tab. 2, VIS-GRPO yields consistent gains across all benchmarks and achieves the best overall performance among all RL algorithms, reaching an average score of 77.1, which is 2.3 points higher than the best-performing baseline. Pixel Reasoner improves over GRPO on several benchmarks by introducing reward heuristics to encourage zoom-based trajectories, but it remains consistently behind VIS-GRPO, suggesting that heuristic encouragement does not reliably produce effective multi-zoom active perception strategies. Depth-Stratified GRPO serves as a one-axis ablation that normalizes advantages only among rollouts with the same zoom depth. It provides a modest gain over GRPO, supporting our analysis that the zoom-depth heterogeneity should be handled. However, it remains clearly behind VIS-GRPO, indicating that pixel-space reasoning requires controlling an additional visual-intent heterogeneity. By stratifying rollouts along both zoom depth and visual intent, VIS-GRPO enables fairer group-relative comparisons and consistently yields the strongest performance.

Qualitative Case Study. Fig. 3 shows an 8K high-resolution, visually cluttered example and compares our model with a representative pixel-space reasoning baseline, Pixel Reasoner [24]. Pixel Reasoner commits early to an incorrect hypothesis: it zooms into a visually plausible but irrelevant region, where the target figurine is in fact absent, and subsequently produces an incorrect answer due to spurious visual cues and hallucinated evidence. In contrast, our model follows a more reliable active-perception routine. It first zooms into a candidate shelf area, correctly concludes that the figurine is not present in the initial crop, and

then performs additional exploratory zooms over alternative candidate regions. It ultimately localizes the figurine and produces the correct answer, illustrating stronger multi-step evidence acquisition ability for pixel-space reasoning. We attribute this improvement to VIS-GRPO's visual-intent stratified grouping, which encourages the learning of exploratory multi-zoom perception strategies.

Training Dynamics of Zoom Usage and Visual Laziness. To diagnose whether an RL algorithm genuinely learns **active perception**, we track the average number of zoom-in operations per sampled trajectory throughout training in Fig. 4. Standard GRPO exhibits a clear zoom-collapse trend: the average number of zoom-ins steadily decreases over training, consistent with the **visual laziness** failure mode, where the model increasingly prefers low-zoom behaviors. Depth-Stratified GRPO partially mitigates early instability by preventing cross-depth comparisons, but its zoom usage still drifts downward in later training, indicating that controlling only the zoom depth is insufficient for learning effective multi-step evidence gathering. Pixel Reasoner maintains roughly one zoom per trajectory, due to its explicit zoom-encouraging reward heuristics; however, encouraging the number of zoom actions alone does not ensure that the learned multi-zoom behaviors are strategically effective. Accordingly, Pixel Reasoner's increase in zoom usage does not consistently translate into higher benchmark performance, as shown in Tab. 2. In contrast, VIS-GRPO steadily increases zoom usage and stabilizes at a much higher level, suggesting that visual-intent stratification provides a more reliable learning signal for exploratory multi-zoom trajectories and prevents the policy from collapsing to zoom-averse solutions.

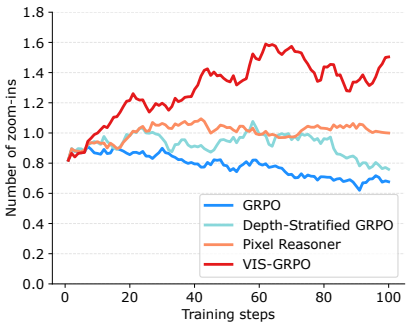
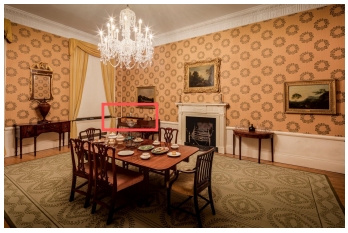


Fig. 4: Training dynamic of zoom-in numbers over training steps.

Effectiveness of Visual-Intent Stratification. To understand why depth-only stratification still struggles to learn reliable multi-zoom strategies, we further analyze the test cases on MME-RealWorld where the model uses more than one zoom. We split these multi-zoom trajectories into `explore` (searching diverse regions) and `verify` (verifying overlapping regions) based on spatial coherence illustrated in Sec. 3.3, and report the visual intent distribution and the corresponding accuracy in Fig. 5.

Depth-Stratified GRPO exhibits a highly imbalanced intent distribution (approximately 9:1) and a noticeable performance gap between `explore` and `verify` (41.6% vs. 28.0%). This is consistent with our analysis that controlling only for zoom depth does not resolve the remaining heterogeneity within the multi-zoom regime: `explore` and `verify` serve different purposes and exhibit different failure modes, yet they are still compared against each other through group-relative normalization. In particular, `verify` requires accurate local verification across overlapping regions and is typically harder to execute early in training; when



Question: What's the color of the uniform of the figurine on the shelf?
Correct Answer: White

Pixel Reasoner

Think Step 1:
The image shows a formal dining room with ornate decor, including a chandelier, patterned wallpaper, and framed artwork. On the left side of the image, there is a shelf with a figurine placed on it.

Now I will zoom in to look clearer at the figurine on the shelf.

Think Step 2:
The cropped image focuses on a decorative wall piece with a figurine mounted above it. The figurine appears to be part of an ornate design, but its details are not clearly visible due to the angle and lighting. ✗

Ours

Think Step 1:
To determine the color of the uniform of the figurine on the shelf, I need to get a closer look at the figurine. The current view is too distant to clearly identify the color of the uniform.

I will zoom in on the shelf area to examine the figurine in detail

Think Step 2:
After examining the cropped image, I can see a dark wooden console table against the wall. On top of this table, there is a large, dark wooden urn with a decorative lid. There is no figurine visible on this shelf. I need to look elsewhere in the room to find the figurine.

I will now scan the rest of the room to locate the figurine.

Think Step 3:
Based on the provided images, I can now clearly see the figurine in question. It is a small, white figurine of a person, standing on top of a wooden piano. The figurine is dressed in a uniform that is predominantly white, with some darker details on the hat and possibly the trim.

Therefore, the color of the uniform of the figurine on the shelf is white. ✓

Fig. 3: Qualitative comparison between Pixel Reasoner and our model on high-resolution, visually complex images. Pixel Reasoner zooms into an irrelevant region and produces a hallucinated answer. In contrast, our model scans multiple locations, correctly localizing the target evidence, and answering accurately.

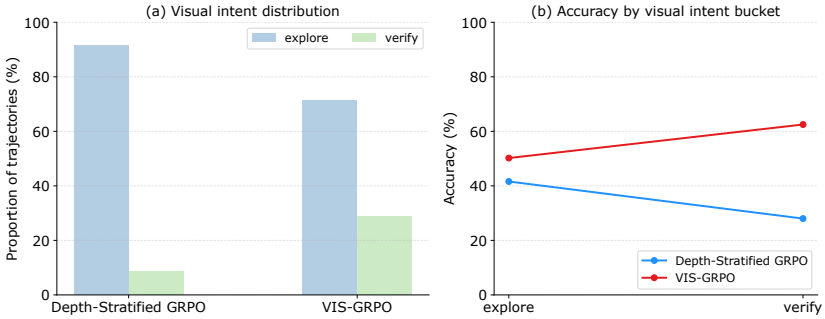


Fig. 5: Visual-intent breakdown of multi-zoom trajectories on MME-RealWorld. (a) Distribution of visual intent for Depth-Stratified GRPO vs. VIS-GRPO; bar heights indicate the fraction of trajectories classified as `explore` or `verify`. (b) Accuracy within each visual-intent bucket.

mixed with `explore` rollouts in the same depth group, `verify` trajectories are more likely to receive negative relative advantages and become under-trained, preventing the emergence of reliable `verify` behavior.

In contrast, VIS-GRPO substantially increases the prevalence of `verify` trajectories (yielding a much more balanced ratio of 7:3), while improving the accuracy of both intents. This aligns with our method design: by stratifying rollouts jointly by zoom depth and visual intent, VIS-GRPO ensures that advantages are computed among strategically comparable trajectories, stabilizing learning signals for intent-specific multi-zoom behaviors and enabling the model to reliably learn effective multi-zoom active perception strategies.

We further include a sensitivity analysis of the spatial-coherence statistic threshold β used for visual-intent stratified grouping in the supplementary material, demonstrating that VIS-GRPO remains robust across a reasonable range of β values.

5 Conclusion

In this paper, we study why reinforcement learning algorithms like GRPO often fail to realize active perception for pixel-space reasoning. We identify a common failure mode, visual laziness, where models trained with standard GRPO increasingly avoid zoom-in actions and collapse toward low-zoom behaviors. Our analysis attributes this behavior to a structural mismatch in group-relative advantage normalization: GRPO compares rollouts that follow fundamentally different perception strategies, which systematically suppresses exploratory multi-zoom trajectories before they can be reliably learned. To resolve this mismatch, we propose Visual-Intent Stratified GRPO (VIS-GRPO), a drop-in replacement for GRPO that aligns advantage estimation with the intrinsic structure of pixel-space reasoning. VIS-GRPO stratifies rollouts jointly by zoom depth and visual intent, and computes normalized advantages within each depth–intent stratum.

This strategy-aligned normalization stabilizes learning signals for intent-specific multi-step zoom behaviors, enabling the emergence of effective multi-zoom perception strategies. Extensive experiments show that VIS-GRPO consistently promotes effective multi-zoom active perception and improves performance across diverse challenging visual understanding benchmarks.

Acknowledgements

This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

References

1. Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., Hooker, S.: Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12248–12267 (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv: 2511.21631 (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:2502.13923 (2025)
4. Chen, H., Tu, H., Wang, F., Liu, H., Tang, X., Du, X., Zhou, Y., Xie, C.: SFT or RL? an early investigation into training R1-like reasoning large vision-language models. Transactions on Machine Learning Research (2025)
5. Chen, X., Zhu, M., Liu, S., Wu, X., Xu, X., Liu, Y., Bai, X., Zhao, H.: Mico: Multi-image contrast for reinforcement visual reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261 (2025)
7. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian,

N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv: 2501.12948 (2025)

8. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: OpenVL-Thinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv:2503.17352 (2025)
9. 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)
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in MLLMs. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
11. Hu, Y., Stretcu, O., Lu, C.T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., Fuxman, A.: Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9590–9601 (2024)
12. Huang, W., Jia, B., Cao, S., Ye, Z., zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. In: The Fourteenth International Conference on Learning Representations (2026)
13. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research **4**, 237–285 (1996)
14. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In: The Fourteenth International Conference on Learning Representations (2026)
15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. Transactions on Machine Learning Research (2025)
16. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: DeepSeek-VL: towards real-world vision-language understanding. arXiv:2403.05525 (2024)
17. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: InfographicVQA. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1697–1706 (January 2022)
18. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Shi, B., Wang, W., He, J., Zhang, K., et al.: MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. arXiv:2503.07365 (2025)

19. OpenAI: 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:2410.21276 (2024)
20. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
21. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36**, 53728–53741 (2023)
22. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)
23. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: DeepSeekMath: pushing the limits of mathematical reasoning in open language models. arXiv: 2402.03300 (2024)
24. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. Adaptive Computation and Machine Learning Series, MIT Press, Cambridge, MA, 2 edn. (2018)
26. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv:2312.11805 (2023)
27. 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:2403.05530 (2024)
28. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv:2503.19786 (2025)
29. 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 Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25 (2025)
30. Wang, X., Song, D., Chen, S., Chen, J., Cai, Z., Zhang, C., Sun, L., Wang, B.: LongLLaVA: Scaling multi-modal LLMs to 1000 images efficiently via a hybrid architecture. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 21419–21436. Suzhou, China (Nov 2025)
31. Wang, X., Yang, Z., Feng, C., Lu, H., Li, L., Lin, C.C., Lin, K., Huang, F., Wang, L.: SoTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
32. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**(3-4), 229–256 (1992)
33. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. arXiv:2405.09711 (2024)

34. 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)
35. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: LLaVA-CoT: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2087–2098 (October 2025)
36. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Wu, Y., Wang, M.: DAPO: An open-source LLM reinforcement learning system at scale. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
37. Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., MA, Z., He, J.: SimpleRL-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In: Second Conference on Language Modeling (2025)
38. Zhang, Y., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R.: MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In: The Thirteenth International Conference on Learning Representations (2025)
39. Zheng, J., Li, J., Cheng, S., Zheng, Y., Li, J., Liu, J., Liu, Y., Liu, J., Zhan, X.: Instruction-guided visual masking. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
40. Zheng, J., Shen, J., Yao, Y., Wang, M., Yang, Y., Wang, D., Liu, T.: Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In: The Thirteenth International Conference on Learning Representations (2025)
41. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In: The Fourteenth International Conference on Learning Representations (2026)
42. Zhu, M., Chen, X., Wang, Z., Yu, B., Zhao, H., Jia, J.: Enhancing LLM knowledge learning through generalization. In: Findings of the Association for Computational Linguistics: EMNLP 2025 (2025)
43. Zhu, M., Chen, X., Wang, Z., Yu, B., Zhao, H., Jia, J.: TGDPO: Harnessing token-level reward guidance for enhancing direct preference optimization. In: Forty-second International Conference on Machine Learning (2025)
44. Zhu, M., Chen, X., Yu, B., Zhao, H., Jia, J.: From noisy traces to stable gradients: Bias-variance optimized preference optimization for aligning large reasoning models. arXiv: 2510.05095 (2025)
45. Zhu, M., Chen, X., Yu, B., Zhao, H., Jia, J.: Stratified GRPO: Handling structural heterogeneity in reinforcement learning of LLM search agents. In: Forty-third International Conference on Machine Learning (2026)