Diversity Over Frequency: Rethinking Tool Use in Visual Chain-of-Thought Agents

Anonymous Authors[1]

Abstract

Visual agents employ tools such as zoom-in cropping within visual chains of thought to access fine-grained details. Prior work has primarily demonstrated the effectiveness of these tools on visual search tasks, leaving their applicability to more diverse and complex visual problems underexplored. In this paper, we move beyond visual search and study challenging visual tasks that require advanced spatial understanding and reasoning, such as 3D spatial reasoning, where agents must not only crop or zoom in on relevant regions but also understand how these local details relate to the global context. We identify a *tool-use collapse phenomenon*: models progressively stop using tools while still achieving higher task accuracy. Moreover, we observe a clear asymmetry: (i) completely eliminating tool use degrades performance, whereas (ii) incentivizing tool use yields only marginal gains despite substantially increasing usage. We find that vanilla training and rewarding tool-use encouragement reduce rollout diversity during training, explaining why higher tool-use does not yield stronger reasoning performance. Motivated by these findings, we encourage diverse rollout exploration by adding an entropy-regularization term to the reinforcement learning objective, which results in the best performance despite tool usage gradually declining during training. Overall, our findings suggest a training-time view of tools as scaffolding, where broader exploration in both text and vision shapes representations that improve despite tool-use collapse.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Recent visual agents increasingly adopt a visual chain-of-thought where visual tools such as zoom-in cropping and grounding serve as intermediate reasoning steps to capture fine-grained evidence. In benchmarks like V* (Wu & Xie, 2024), employing such tool-mediated reasoning chains has been shown to be critical for solving high-resolution tasks (Zheng et al., 2025; Zhang et al., 2025a; Lai et al., 2025). This formulation frames the iterative execution of tools as the reasoning process itself, a strategy that is further exemplified by practical systems like ChatGPT-o3 (OpenAI, 2025). Consequently, there is growing interest in training agents to actively leverage these tool-based reasoning paths for broader visual perception challenges (Shao et al., 2024; Lai et al., 2025; Zheng et al., 2025; Wang et al., 2025a; Wu et al., 2025; Su et al., 2025; Zhang et al., 2025a).

However, it remains unclear whether such tool-driven gains generalize to broader, more diverse, and more complex visual tasks. Most prior studies focus on visual search scenarios where simple zooming or cropping of relevant regions is sufficient to solve the task. In contrast, we investigate visual tasks that demand more advanced reasoning, using 3D spatial reasoning from 2D images as a representative testbed for general spatial understanding. Unlike visual search tasks that can be solved through simple detection, spatial reasoning demands the identification of implicitly relevant regions and an understanding of their relationships to the global spatial structure. Thus, this task requires agents to learn selective tool use, reasoning about when additional visual evidence is beneficial for complex spatial understanding. Studying this regime is crucial because, as we show, naive application of reinforcement learning (RL) struggles to regulate selective tool use in complex spatial reasoning tasks, leading to unexpected failure modes.

Our first finding is a consistent training pathology we call tool-use collapse. Starting from a pretrained tool-using agent, vanilla reinforcement learning fine-tuning (RFT) improves accuracy while driving the tool-use rate, defined as the fraction of rollouts that invoke a tool, to near zero. This collapse occurs even when employing methods designed to mitigate length penalties associated with multi-step tool-based reasoning (Lai et al., 2025). In contrast, tool encour-

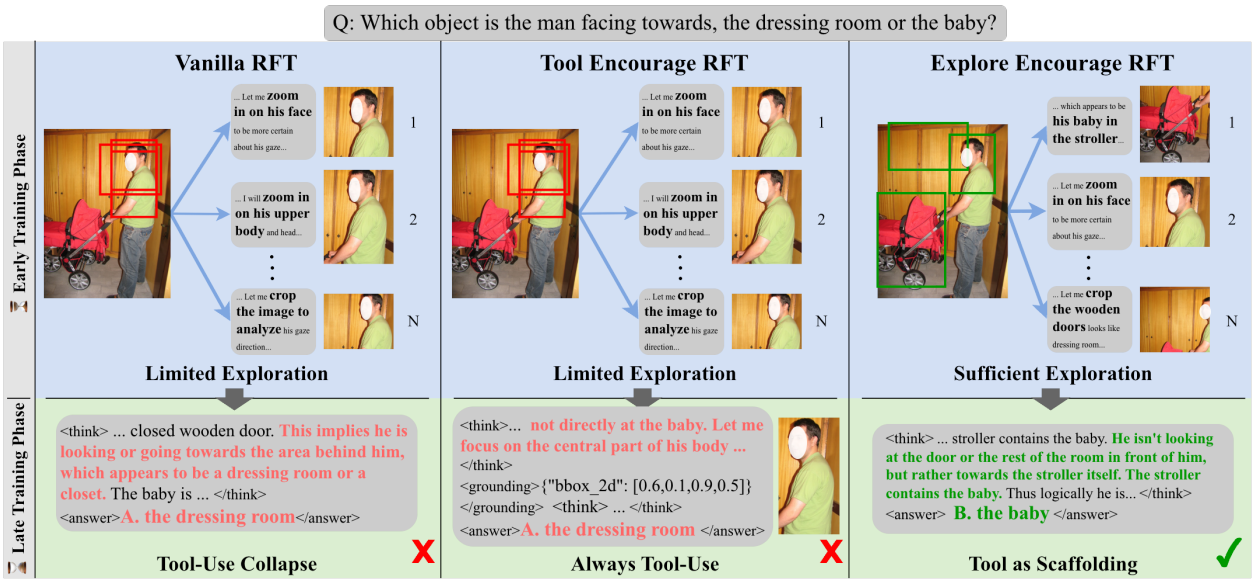


Figure 1. **Effect of Early Exploration on Late Training Phase** This figure compares agent behaviors in Early Training (top) and Late Training (bottom) phases. Both vanilla (Left) and tool-encouraged RL (Middle) suffer from limited early exploration in both text and visual modalities. This lack of rollout diversity restricts models to local optima with marginal gains regardless of tool usage frequency. In contrast, encouraging exploration (Right) leads to better performance even when the agent rarely uses tools. We call this the scaffolding effect where early tool use acts like temporary supports to build reasoning capabilities that remain effective even after tool use is discarded.

aging reward design (Zheng et al., 2025; Wang et al., 2025a) in RFT consistently uses tools but yields only marginal accuracy gains. Fig. 1 illustrates this asymmetry: vanilla RFT (left) and tool-encouraging RFT (middle) show that both fine-tuning approaches suffer from limited early exploration, repetitively fixating on salient features (e.g., the man's face). Conversely, the right column demonstrates that fostering diverse exploration acts as *scaffolding* by covering broader contexts (e.g., the stroller and wooden door), enabling the accurate inference. To make this mechanism explicit, we quantify exploration in both text and vision, which reveals that vanilla RFT and tool-use-focused RFT lead to diversity degradation in both modalities. This confirms that the scaffolding effect relies not on tool frequency, but on preventing premature convergence to repetitive reasoning and visual fixation.

Motivated by these findings, we employ adaptive entropy regularization (Zhang et al., 2025b) during RFT to encourage such diverse exploration, resulting in the best performance despite tool usage gradually declining during training. Overall, our results support a training-time view of tools as scaffolding, where early tool-mediated exploration shapes representations that enable accurate tool-less inference.

In summary, our contributions are as follows:

1. We identify tool-use collapse under vanilla RFT of a pretrained tool-using visual agent for visual reasoning tasks (e.g., 3D spatial reasoning), where accuracy improves while tool use almost vanishes.
2. We show a clear asymmetry: removing tools degrades performance, whereas reward-based tool encouragement while RFT drives constant tool calls with only marginal gains. We link both outcomes to insufficient exploration measured in both textual reasoning and visual tool behavior.
3. We show that encouraging exploration by adding an adaptive entropy regularization term increases rollout diversity during training and achieves the best performance while tool usage still naturally declines, supporting the view of tools as scaffolding during training.

2. Preliminaries & Experimental Setup

2.1. Base Model and Tool Interface

For all experiments, we initialize from the pretrained weights of Mini-o3 (Lai et al., 2025). Mini-o3 is a `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025)-based VLM that has been trained with supervised fine-turning (SFT) and RFT to perform iterative visual tool use (e.g., zoom-in cropping) via a structured interaction format.

Agentic Interaction Loop Given a spatial reasoning text query q and input image I_0, the policy model π_θ generates a rollout τ through a thought–action–observation loop. A rollout τ comprises a single text-only step or a multi-turn sequence of generated `<think>` spans, optional `<grounding>` actions, and resulting observations until the model outputs a terminal `<answer>`.

At step i, the model produces:

- **Thought** T_i: Internal reasoning conditioned on the interaction history $h_{<i} = \{(T_j, A_j, O_j)\}_{j<i}$ and current observation.
- **Action** A_i: Either emit a `<grounding>` action to request a crop, or terminate with a final `<answer>`.
- **Observation** O_i: The resulting image patch or terminal state.

The rollout terminates when the model outputs a final answer or interaction turns. In practice, rollouts follow explicit tags, including an initial `<think>` span, an optional `<grounding>` action with parameters, and a terminal `<answer>` span. See Fig. C for rollout examples.

A `<grounding>` action is parameterized by two components:

$$A_i^{\texttt{<grounding>}} = (\texttt{bbox_2d}, \texttt{source}), \tag{1}$$

where $\texttt{bbox_2d} \in [0,1]^4$ specifies normalized coordinates (x_1, y_1, x_2, y_2) for the zoom-in region, and $\texttt{source} \in \{I_0, O_1, \ldots, O_{i-1}\}$ indicates the image to crop from, either the original input or any prior observation. This design allows iterative refinement by zooming into previously cropped regions.

Observation Generation Executing $A_i^{\texttt{<grounding>}}$ produces observation O_i by cropping the specified bounding box from the source image. The cropped patch is appended to the conversation history, enabling the model to reason over both the initial global observation (from I_0) and fine-grained local evidence (from $\{O_j\}$).

Tool-Based vs. Tool-Free Rollouts. A rollout τ may invoke `<grounding>` multiple times (tool-based) or proceed directly to `<answer>` using only the initial global observation I_0 (tool-free). We quantify tool utilization using the Tool Use Ratio, defined as the fraction of rollouts containing at least one `<grounding>` action. Tool-based rollouts are typically longer in token count and interaction turns, as each `<grounding>` action introduces additional image tokens and reasoning steps. This asymmetry becomes critical during RFT, where optimization pressure can favor shorter, lower-variance paths, a phenomenon we analyze in Sec. 3.1.

2.2. Task and Datasets

Training Data We train on 1.2k 3D spatial reasoning samples derived from SpatialReasoner (Ma et al., 2025b), formatted as (image, question, answer) triplets. Questions require understanding of 3D spatial relationships such as relative height, depth ordering, and orientation that cannot be reliably inferred from 2D image coordinates alone. Images are sourced from OpenImages (Kuznetsova et al., 2020) and depict diverse real-world scenes with multiple objects. Critically, questions often reference objects that are small or ambiguous in the global view, creating opportunities where tool-based zoom-in actions could aid spatial reasoning.

Evaluation Benchmarks: 3DSRBench We select 3DSRBench (Ma et al., 2025a) as our primary testbed because it presents a mixed-utility environment essential for studying selective tool use. Unlike visual search tasks dominated by small targets, 3DSRBench covers a broad spectrum of object scales (see Appendix C.2 for a detailed comparison), requiring agents to distinguish when tools are truly necessary. The benchmark contains 5,250 multiple-choice questions on MS-COCO images across four types: Height, Location, Orientation, and Multi-object. Questions avoid 2D shortcuts (e.g., camera pitch variations decorrelate vertical position in image vs. 3D height). Model outputs can be free-form (e.g., `C`, `the answer is C`, or `C. ...`); we therefore compute accuracy using a fixed VLM-as-judge (`Qwen2.5-VL-7B-Instruct`) that normalizes each response and determines which option it supports, then compares it against the ground-truth option. All evaluation is reported as Avg@8, where we sample 8 independent generations per question and average the resulting accuracy.

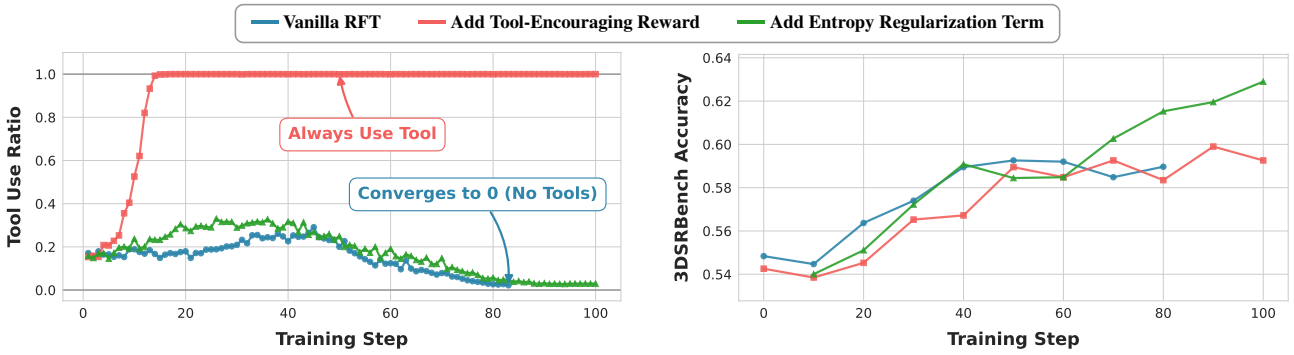
2.3. Reinforcement Learning Objective

We optimize the policy π_θ via reinforcement learning to maximize the expected reward on spatial reasoning tasks:

$$\mathcal{J}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \tau\sim\pi_\theta(\cdot|q,I_0)}\big[R(\tau, q)\big], \tag{2}$$

where q denotes a question sampled from the training distribution $\mathcal{D}$, τ is a sampled rollout, and $R(\tau, q)$ evaluates the final correctness of the terminal response in τ for q.

Following Mini-o3 (Lai et al., 2025), we adopt DAPO (Yu et al., 2025), a stabilized group-based policy optimization method built on GRPO (Guo et al., 2025) that updates the policy from sampled rollouts using group-normalized advantages, together with stabilization techniques used in Mini-o3 (e.g., clip-higher, dynamic sampling) and a token-level policy loss. In practice, we report results up to 100 training steps: beyond this point, we observe oscillatory dynamics where validation performance no longer improves reliably. We additionally note a practical limitation induced by DAPO's dynamic sampling. When a sampled group yields degenerate rollouts (e.g., all rewards are identical, such as all 0 or all 1), the batch is rejected and rollouts are regenerated from a new batch. If the number of such resampling attempts exceeds 10 within a single update step, the wall-clock cost increases sharply, making continued training inefficient; we therefore terminate runs at this point.



(a) Tool Use Ratio

(b) Evaluation on 3DSRBench

Figure 2. **Analysis of Training Dynamics.** (a) Tool Use Ratio over training defined as the fraction of rollouts that contain at least one grounding action. (b) Validation accuracy trends. *Vanilla RFT* (blue) and *Tool-Encourage reward* (red) exhibit divergent tool usage behaviors despite similar limited accuracy gains. In contrast, adding an *entropy regularization term* (green) encourages exploration which leads to a significant increase in final performance.

Policy Optimization For each question q, we sample a group of G rollouts $\{\tau_i\}_{i=1}^{G}$ from the current policy $\pi_{\theta_{\text{old}}}(\cdot \mid q, I_0)$ and compute rewards. The policy is updated via clipped importance sampling with group-normalized advantages:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\, A_i\Big)\right]. \tag{3}$$

where $\rho_i = \pi_\theta(\tau_i \mid q, I_0)/\pi_{\theta_{\text{old}}}(\tau_i \mid q, I_0)$ and $A_i = (r_i - \mu_r)/\sigma_r$ with group statistics μ_r, σ_r.

To mitigate premature termination bias that would otherwise discourage longer, tool-using rollouts, we adopt over-turn masking (Lai et al., 2025). Rollouts that exceed context length (our setting 6) or turn limits are excluded from advantage computation rather than treated as failures, preventing the policy from being implicitly penalized for invoking `<grounding>`.

3. Collapse, Explosion, and Exploration Control

3.1. Vanilla RFT Leads to Tool-Use Collapse

We first investigate the behavior of visual agents under vanilla RFT conditions to establish a baseline and analyze the necessity of visual tools.

Examine Transferable Agentic Ability We train the Mini-o3 agent using standard GRPO with DAPO stabilization, as described in Section 2.1, on 1.2k spatial reasoning questions. The model is warm-started from a SFT-RFT checkpoint that is already trained to use `<grounding>` actions.

The Phenomenon of Tool-Use Collapse: Accuracy Rises as Tool-Use Vanishes As shown in Figure 2, we observe a distinct training dynamic termed **tool-use collapse**. While the validation accuracy on 3DSRBench improves steadily to reach 59.2% at saturation, the tool use ratio declines monotonically, converging to near 2% by 80 training steps. Notably, the agent maintains a non-trivial tool usage ratio (around ∼20%) during the early stage of training before the collapse accelerates.

Optimization Preference for No Tools We hypothesize that this collapse is driven by an asymmetry in the optimization landscape. A tool-based rollout requires a multi-step interaction loop that generates a `<grounding>` action, processes additional visual tokens from the cropped image, and integrates multi-turn observations. In contrast, tool-free reasoning, producing a direct answer from the global image observation without invoking any tool, follows a shorter rollout. Crucially, this trend still appears even when we attempt to mitigate GRPO's implicit preference for shorter generations by replacing the length-sensitive term with DAPO's token-level policy loss (Yu et al., 2025). While this does not by itself establish a causal mechanism, it suggests that tool-use collapse cannot be explained solely as an artifact of GRPO's length bias.

Finally, to rule out the concern that tool-use collapse is an artifact of training on a small 1.2k dataset, we repeat the same training protocol on the larger SAT dataset (6k samples) (Ray et al., 2024) and observe a similar collapse. Detailed results on SAT training are provided in Appendix B.1.

3.2. Ablation: Are Tools Useful?

The collapse observed in the previous experiment raises a critical question: If the agent discards tools and still achieves high accuracy, are tools genuinely unnecessary? To verify this, we conduct an ablation study where tools are completely prohibited.

Tool-Banned Training and Evaluation We train an otherwise identical agent but enforce a strict ban on tool us-

age throughout both training and evaluation. Specifically, we prevent the model from emitting the tool trigger tags (<grounding>, </grounding>) during rollout generation and do not execute any zoom-in/crop operations. As a result, the agent must rely exclusively on tool-free reasoning from the initial global observation, without receiving any tool-mediated visual inputs at any stage.

Early Tool Experience Matters The tool-banned agent achieves a final accuracy of 58.1% which represents a 1.1% drop compared to vanilla RFT. This reveals a paradox where the vanilla RFT outperforms the strictly tool-free agent despite eventually abandoning tool usage itself.

Tools as Training-Time Scaffolding This result suggests that the early tool experience (approximately ∼20% tool-use ratio in the baseline) was not wasted but served as training-time scaffolding. Even though tool usage later fades, these early tool-augmented rollouts may provide a richer exploration history and learning signals that transfer to the eventual tool-free policy. In contrast, the tool-banned training, lacking this initial tool-mediated exploration, fails to reach the same performance peak.

3.3. Do Incentive Reward Designs Improve Tool Use and Accuracy?

Sec. 3.1 and Sec. 3.2 revealed a paradoxical pattern. Under standard RL fine-tuning, tool usage collapses toward zero while accuracy improves. Yet removing tools entirely from the beginning reduces the final accuracy (58.1% vs. 59.2%). This indicates that tool interactions can benefit learning dynamics even if tools are rarely used at convergence. We therefore ask a sharper question: If early tool exposure matters, should we deliberately increase tool usage to improve performance?

A key premise of our analysis is that tool usage frequency is not the objective.

Tool-Encouraging Reward Design We compare two representative tool incentivization schemes. While Mini-o3 (Lai et al., 2025) encouraged tool usage passively by omitting negative rewards for long multi-turns, we introduce an explicit reward to directly incentivize tool use.

(1) Tool bonus. Following DeepEyes (Zheng et al., 2025), we use a rollout-level reward that combines final correctness with a conditional bonus for successful tool use:

$$\begin{aligned} R_{\text{DE}}(\tau, q) = &\ \mathbb{I}[y(\tau) = y^*(q)] \\ &+ \lambda_{\text{tool}}\, \mathbb{I}[y(\tau) = y^*(q)]\, \mathbb{I}[u(\tau) = 1]\,, \end{aligned} \tag{4}$$

where $y(\tau)$ is the extracted option label using the fixed VLM-as-judge protocol described in Sec. 2.1, $y^*(q)$ is the ground-truth label, and $u(\tau)$ indicates whether the rollout contains at least one <grounding> action.

(2) Curiosity reward. Following PixelReasoner (Wang et al., 2025a), we augment correctness with a curiosity bonus that encourages tool-mediated reasoning when the policy under-adopts it for a given query:

$$\begin{aligned} R_{\text{PR}}(\tau, q) = &\ \mathbb{I}[y(\tau) = y^*(q)] \\ &+ \alpha\, \max\big(H - \text{RaPR}(q),\, 0\big)\, \mathbb{I}[u(\tau) = 1] \\ &+ \beta\, r_{\text{penalty}}(\tau), \end{aligned} \tag{5}$$

where $\text{RaPR}(q)$ denotes the rate of tool-mediated reasoning for query q, following the Rate of Pixel-space Reasoning definition in PixelReasoner. Concretely, we estimate $\text{RaPR}(q) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot|q, I_0)}[u(\tau)]$, which corresponds to the expected frequency of <grounding> usage for query q, and H is the target adoption threshold. Following PixelReasoner, we impose an upper bound on tool calls within a rollout via $r_{\text{penalty}}(\tau) = \min\big(N - n_{\text{tool}}(\tau), 0\big)$, where $n_{\text{tool}}(\tau)$ is the number of tool calls in τ. This penalty is 0 when $n_{\text{tool}}(\tau) \leq N$ and becomes negative only when tool calls exceed N.

For both schemes, we follow the original papers' hyperparameter choices (e.g., λ_{tool}, α, β, H, and N) in our experiments. For clarity, we report the Tool bonus results in the main text. The Curiosity reward shows the same trend and is deferred to Appendix B.2.

Tool Explosion The results are shown in Fig. 2. Under the DeepEyes' tool bonus, the policy rapidly increases its reliance on tool calls. Starting from the same warm-started checkpoint (Tool Use Ratio ∼20%), the ratio rises throughout training and reaches 100% at saturation, meaning that every rollout invokes at least one tool call. Despite this persistent tool usage, the final accuracy on 3DSRBench increases only modestly to 59.9% (vs. 59.2% for the baseline).

Increased tool usage does not necessarily correspond to greater diverse exploration Taken together with the tool-banned ablation in Sec. 3.2, these results indicate that the limitation is not the mere availability of tools. Tools can provide a useful learning signal early in training, yet simply driving tool usage to 100% yields only marginal gains. This motivates a closer diagnosis of what changes in learning dynamics under vanilla RFT and Tool-Encourage RFT.

Specifically, we ask whether these training runs maintain diverse reasoning rollouts as learning progresses. We characterize diversity along two axes that correspond to our tool interface: (i) the textual reasoning diversity that precedes a crop decision, and (ii) the visual regions diversity selected by crop actions.

To quantify diversity in the textual reasoning space, we analyze the reasoning produced before issuing a crop via <grounding>. This choice is motivated by an entropy

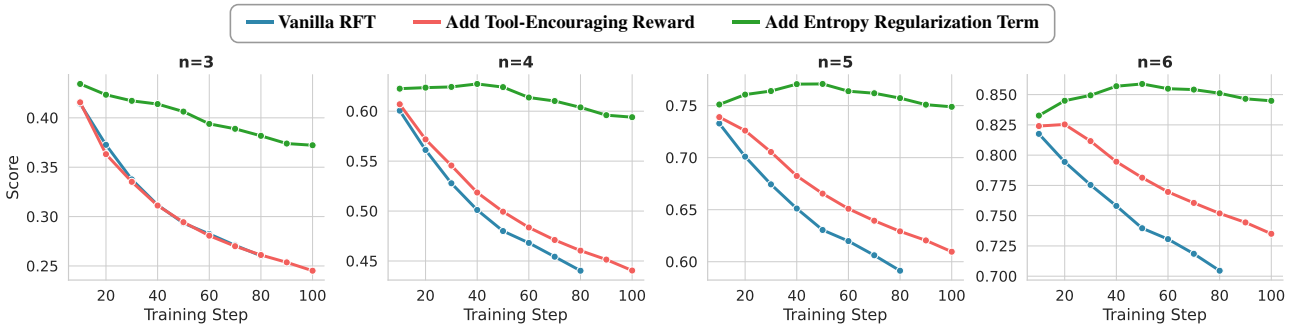


Figure 3. **Quantitative Analysis of Textual Exploration.** We report the distinct n-gram ratio (Li et al., 2016) ($n \in \{3, 4, 5, 6\}$) to measure token diversity of generated rollouts. Both the *Vanilla RFT* (Blue) and *Tool-Encourage* (Red) suffer from diversity degradation, indicating increasingly repetitive reasoning patterns. In contrast, adding an *Entropy regularization term* (Green) maintains higher textual diversity, suggesting that the scaffolding effect relies on preventing premature convergence in the reasoning space.

Method	mIoU (↓)	CLIP (↑)	Visual Behavior
Vanilla RFT	0.554	0.184	High Fixation
Tool-Encourage	0.557	0.187	High Fixation
Entropy-Regularized	0.494	0.184	Active Exploration

Table 1. **Quantitative Analysis of Visual Exploration.** We measure spatial diversity via mIoU computed between crop boxes across rollouts for a single query (lower indicates broader exploration) and semantic relevance via CLIP alignment between the crop and a keyword prompt. We form the keyword prompt from question nouns (the red words in Fig. 4). Higher CLIP indicates better alignment; absolute values are reported for comparison. *Entropy-Regularized* promotes active visual exploration without losing semantic focus, whereas others suffer from spatial fixation.

probe: we found that uncertainty is not primarily expressed in the `<grounding>` box coordinates themselves, but earlier in the preceding textual reasoning, suggesting that the pre-`<grounding>` span often drives where the agent decides to crop. Following the interaction format in Sec. 2.1, each rollout contains an initial `<think>...</think>` span that precedes either a `<grounding>` action (tool-based) or a direct answer (tool-free), enabling consistent extraction of the pre-`<grounding>` reasoning span across methods. As shown in Fig. 3, both vanilla RFT (red) and Tool-Encourage (blue) exhibit a monotonic decrease in n-gram diversity over training, indicating increasingly templated pre-`<grounding>` reasoning.

We also diagnose the visual side of exploration, since higher tool usage does not necessarily imply broader visual evidence. Our goal is to verify, in order, (ii-a) whether the model explores diverse crop locations in image space, and (ii-b) whether the visited crops remain semantically tied to the question rather than drifting to irrelevant regions.

Concretely, we take the checkpoint at step 40, where vanilla RFT training exhibits the most active tool use, and compare it against the Tool-Encourage checkpoint at the same step. For each image–question pair in the 1.2k training set, we sample 50 stochastic rollouts from each policy and collect all crop boxes produced via `<grounding>`. To quantify diversity of crop locations, we compute the mean pairwise IoU between crop boxes generated for the same image–question pair; lower IoU indicates that the policy visits a broader set of regions, whereas higher IoU indicates visual fixation. To quantify semantic relevance, we compute a CLIP score between each cropped region and a keyword extracted from the question (nouns; e.g., the red-highlighted words in Fig. 4), and average scores across rollouts.

Tab. 1 summarizes the results. Tool-Encourage invokes `<grounding>` substantially more often than vanilla training (roughly three times more crops per query), but the crop-location statistics remain similar across the two methods (mean pairwise IoU > 0.55 in both cases). For example, Fig. 4a shows that vanilla RFT concentrates its crops around the man's upper body. Moreover, the semantic relevance of selected crops is comparable (similar CLIP scores), indicating that increased tool use does not translate into broader coverage or more question-aligned visual evidence. These results reinforce that simply encouraging tool invocation is insufficient to expand the exploration history during training, motivating the need to explicitly sustain exploration during RL fine-tuning in the next section.

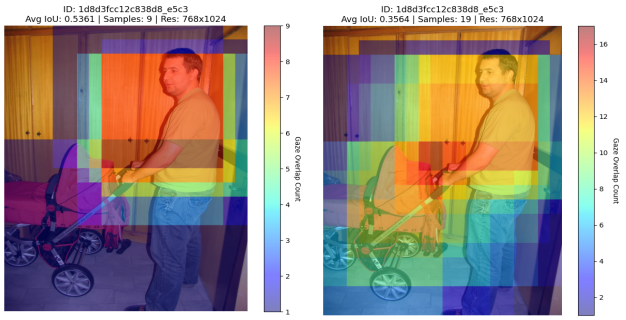
3.4. Does Diversity Matter? Encouraging Exploration via Entropy Regularization

The results in Sec. 3.3 show that increasing tool usage via reward incentives is insufficient to sustain exploration. We therefore treat tools as a means to broaden exploration history during training and use entropy regularization as a simple control knob to prevent premature convergence.

Entropy Regularization as an Exploration Control Knob We augment the GRPO objective with a policy-entropy regularization term:

$$\mathcal{J}_{\text{ent}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) + \lambda_t \cdot \mathbb{E}_{q,\tau}\left[\bar{\mathcal{H}}(\tau)\right], \qquad (6)$$

where $\lambda_t \geq 0$ is the entropy coefficient at update step t and $\bar{\mathcal{H}}(\tau)$ is the mean token-level entropy over the entire rollout τ. For an autoregressive policy, token-level entropy



(a) **Vanilla RFT** *(b)* **Entropy-Regularized**

Q: *Which object is the man facing towards, the dressing room or the baby?*

Figure 4. **Visual Exploration During Training.** Comparison of crop distributions at an early-training checkpoint where tool use still actively explores. While the baseline **(a)** narrowly focuses on the salient subject, the entropy-regularized model **(b)** actively explores contextually relevant regions (the stroller) referenced in the query, demonstrating wider exploration.

at generation step k with state s_k is

$$\mathcal{H}(\pi_\theta(\cdot \mid s_k)) = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid s_k) \log \pi_\theta(v \mid s_k), \quad (7)$$

and we compute $\bar{\mathcal{H}}(\tau)$ by averaging $\mathcal{H}(\pi_\theta(\cdot \mid s_k))$ across all generated tokens in the rollout.

In preliminary experiments, we found a fixed entropy coefficient to be brittle. When it is too small, it has negligible effect on exploration, whereas larger values can destabilize training and lead to degenerate generations, such as mixed-language outputs or repetitive token loops. To avoid manual tuning, we adopt an adaptive entropy coefficient using proportional feedback, following prior work (Zhang et al., 2025b):

$$\lambda_t = K_p \left[\mathcal{H}_{\text{target}} - \mathcal{H}_t\right]_+, \quad (8)$$

where $\mathcal{H}_t$ is the current batch entropy measured over full rollouts, $\mathcal{H}_{\text{target}} = 0.9$ is the target entropy, $K_p = 0.03$ is the gain, and $[\cdot]_+ = \max(\cdot, 0)$. This rule increases exploration pressure only when entropy drops below the target and otherwise sets λ_t to zero.

Entropy Regularization Improves Best Accuracy. With entropy regularization, the agent reaches a best validation accuracy of 62.9% on 3DSRBench during training. Although the run starts from the same warm-started checkpoint (Tool Use Ratio ∼20%), the Tool Use Ratio decreases over training and is 3% at the end of training (measured on training-batch rollouts). We report this run as the green curve in Fig. 2 for consistency with Fig. 3.

Explore More, Tools Fade, Performance Persists. This training also uses tools only rarely by the end of training,

Method	Tools?	Acc.	Tool Usage (Init→Sat)
vanilla RFT	Yes	59.2%	∼20% → ≈2%
Tool-banned	No	58.1%	0% → 0%
Reward-encouraging reward design	Yes	59.9%	∼20% → 100%
Entropy-regularized	Yes	**62.9%**	∼20% → ≈3%
Tool-banned & Entropy-regularized	No	57.8%	0% → 0%

Table 2. **Experiment Summary.** Vanilla RFT collapses tool use (20%→2%), banning tools hurts accuracy, reward incentives saturate tool use (20%→100%) with marginal gains, and entropy regularization yields the best accuracy with low tool use at saturation.

yet it substantially outperforms both the vanilla RFT and the reward-incentivized setting. Notably, directly incentivizing tool invocation drives near-constant `<grounding>` calls but yields only marginal gains, which makes it difficult to attribute the improvement to learning to use tools more at inference time. Instead, the evidence points to a training-time effect: what matters is not tool frequency itself, but whether training sustains exploration in the reasoning process and the visual evidence it visits.

This interpretation is supported by our exploration diagnostics in both text and vision. In the textual reasoning space, Fig. 3 green line shows that the entropy-regularized run maintains higher n-gram diversity, whereas both vanilla RFT and Tool-Encourage become increasingly templated in the pre-`<grounding>` `<think>` span. In the visual tool space, Tab. 1 shows that the entropy-regularized policy explores more diverse crop locations (mIoU 0.494 vs. 0.554/0.557 for Vanilla/Tool-Encourage) while preserving comparable semantic relevance (CLIP 0.184 vs. 0.184/0.187), and Fig. 4b provides a qualitative example of this broader coverage. Taken together, these results suggest that the key benefit comes from a richer exploration history during training. As in vanilla RFT, tool usage still fades late in training, but the resulting tool-free policy is stronger because it was shaped by more diverse reasoning rollouts and visual evidence earlier on.

At the same time, this result raises an important follow-up question. Is the improvement driven by a general effect of maintaining higher policy entropy, or does it specifically depend on tool-mediated visual exploration during training? To disentangle these factors, we next compare entropy regularization with and without tool access.

3.5. Ablation: Is the Entropy Gain Tool-Dependent?

Tool-Banned Entropy Regularization. To test whether entropy regularization yields tool-agnostic gains, we repeat the tool-banned protocol in Sec. 3.2 while adding the same entropy control as in Section 3.4.

Entropy Gains Require Tool Access. Under strict tool banning, entropy regularization does not improve performance. The tool-banned + entropy run reaches a best validation accuracy of 57.8%, which is not higher than the tool-banned baseline and remains far below the tool-enabled

Method	3DSRBench (Acc, %)	CV-Bench-3D (Acc, %)
Generalist Baseline		
Qwen2.5-VL 7B (Bai et al., 2025)*	48.4	**82.9**
Visual Agents (Zero-shot)		
DeepEyes (Zheng et al., 2025)	51.6	76.7
Mini-o3 (Lai et al., 2025)	**54.5**	77.6
Specialist Reference		
SpatialReasoner (Ma et al., 2025b)*	60.3	80.3
Ours Experiments (RL Fine-tuning)		
vanilla RFT (Base)	59.2	76.7
w/ Tool-Encourage Reward	59.9	74.5
w/ Entropy-Regularization	**62.9**	78.8

Table 3. **Impact of Exploration on Generalization.** We compare performance on 3DSRBench and CV-Bench-3D (Tong et al., 2024). While vanilla RFT and tool-encourage rewards degrade performance on the general benchmark compared to the pre-trained Mini-o3 weight, Entropy-Regularization is the only method that improves both specialized and general capabilities, confirming that diverse exploration shapes more robust representations. (* are from SpatialReasoner (Ma et al., 2025b))

entropy-regularized setting (62.9%). This contrast indicates that the gains from entropy regularization are not purely tool-agnostic, but depend on tool-mediated visual exploration during training. When tools are unavailable, increasing exploration pressure can be neutral or even detrimental, suggesting that the scaffolding benefit arises from exploring additional visual evidence rather than from entropy alone.

4. Discussion

4.1. Generalization Analysis on CV-Bench-3D

To verify whether adding the entropy-regularization term yields genuine improvements on spatial understanding rather than merely overfitting to the target task, we extended our evaluation to CV-Bench-3D (Tong et al., 2024). This benchmark consists of 1,200 questions that evaluate fundamental depth perception capabilities, including depth order and relative distance estimation from monocular images.

As presented in Tab. 3, Vanilla RFT exhibits a notable trade-off between task-specific optimization and general visual understanding. Comparing against the pre-trained Mini-o3 accuracy of 77.6%, the standard vanilla RFT results in a slight regression to 76.7%. The performance degradation is even more pronounced when explicitly forcing tool use, where the accuracy drops significantly to 74.5%. These results imply that restricting agent behavior through either premature collapse or forced fixation could potentially compromise generalist capabilities.

In contrast, the inclusion of an entropy-regularization term yields the only approach that surpasses the base model, achieving 78.8% accuracy. We interpret this improvement as empirical evidence consistent with the hypothesis that diverse exploration during training may serve as a form of visual scaffolding. By encouraging the agent to broadly engage with various visual cues, the model appears to acquire more robust spatial understandings that possess better transferability to general depth tasks, rather than strictly overfitting to the tool-use patterns of the training dataset. Overall, these findings underscore the potential importance of exploration diversity for enhancing both specific reasoning capabilities and general visual understanding.

4.2. Do Visual Agents Already Have Spatial Reasoning Ability?

Partially Yes Tab. 3 suggests that tool-using visual agents already exhibit non-trivial spatial reasoning ability even without any explicit spatial training. Compared to the generalist baseline (Qwen2.5-VL 7B), both DeepEyes and Mini-o3 improve accuracy on 3DSRBench, which indicates that agentic tool use and visual chain-of-thought can recover fine-grained evidence through tool-mediated inspection that is sometimes missed from a single global view. At the same time, this advantage does not transfer uniformly. On CV-Bench-3D, DeepEyes underperforms the generalist baseline, and Mini-o3 shows a smaller margin, which highlights that tool use is not a guaranteed fix for spatial reasoning and that different spatial benchmarks can stress different failure modes.

A natural concern is that our gains could simply reflect starting from a strong spatial agent such as Mini-o3. The specialist reference SpatialReasoner provides a useful comparison here. SpatialReasoner reaches 60.3% on 3DSRBench by explicitly incorporating 3D coordinates into textual reasoning during its SFT-RFT pipeline. Our best model achieves 62.9% on 3DSRBench without using such explicit 3D representations or any dedicated spatial supervision in the initialization. This indicates that the main driver of improvement is not an unusually strong spatial pretraining recipe, but the training-time exploration dynamics studied in this work.

5. Conclusion

We investigated the role of visual tools in an "optional tool" regime using 3D spatial reasoning as a testbed. We identified a consistent *tool-use collapse* dynamic where tool use drops to near zero as accuracy improves, even under methods designed to mitigate length penalties. This collapse does not imply tools are unnecessary; banning them degrades performance, confirming that early interactions provide essential learning signals. However, forcing usage via rewards yields only marginal gains and homogenizes reasoning rollouts. Adaptive entropy regularization proves more effective, achieving the best performance while allowing tool use to fade naturally. Our quantitative analysis confirms that this approach prevents homogenized rollouts, sustaining broader exploration in both textual reasoning and visual crop selection. These findings support a view of tools as *scaffolding*. Early tool access facilitates broad exploration of evidence during training, shaping robust representations that enable accurate inference even after explicit dependence fades.

Impact Statement

This work provides insights into the dynamics of tool use in visual agents, suggesting that diverse exploration during training may serve as effective scaffolding for robust representation learning. By reducing dependency on inference-time tools, our findings could contribute to the development of more efficient autonomous systems, though further research is warranted to ensure these internalized representations remain reliable across varied real-world contexts.

References

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.

Du, Y., Zhou, K., Min, Y., Ling, Y., Zhao, W. X., and Wu, Y. Revisiting the necessity of lengthy chain-of-thought in vision-centric reasoning generalization. *arXiv preprint arXiv:2511.22586*, 2025.

Fu, X., Liu, M., Yang, Z., Corring, J. R., Lu, Y., Yang, J., Roth, D., Florencio, D., and Zhang, C. Refocus: Visual editing as a chain of thought for structured image understanding. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=a7qFlPOTix.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Krishna, R. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Lai, X., Li, J., Li, W., Liu, T., Li, T., and Zhao, H. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, W. B. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 110–119, 2016.

Ma, W., Chen, H., Zhang, G., Chou, Y.-C., Chen, J., de Melo, C., and Yuille, A. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025a.

Ma, W., Chou, Y.-C., Liu, Q., Wang, X., de Melo, C. M., Xie, J., and Yuille, A. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=hFaXVjRFHI.

OpenAI. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, April 2025. Accessed: 2026-01-12.

Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B. A., Krishna, R., et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., and Chengm, Y. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

Tong, P., Brown, E., Wu, P., Woo, S., IYER, A. J. V., Akula, S. C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Wang, H., Su, A., Ren, W., Lin, F., and Chen, W. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Wang, Y., Chen, Q., Li, Z., Wang, S., Guo, S., Zhang, Z., and Wei, Z. Simple o3: Towards interleaved vision-language reasoning. *arXiv preprint arXiv:2508.12109*, 2025b.

Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., and Nahrstedt, K. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025.

Wu, P. and 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.

Xu, C., Wang, Y., Cai, Y., Hooi, B., and Li, S. Visual cot makes vlms smarter but more fragile. *arXiv preprint arXiv:2509.23789*, 2025.

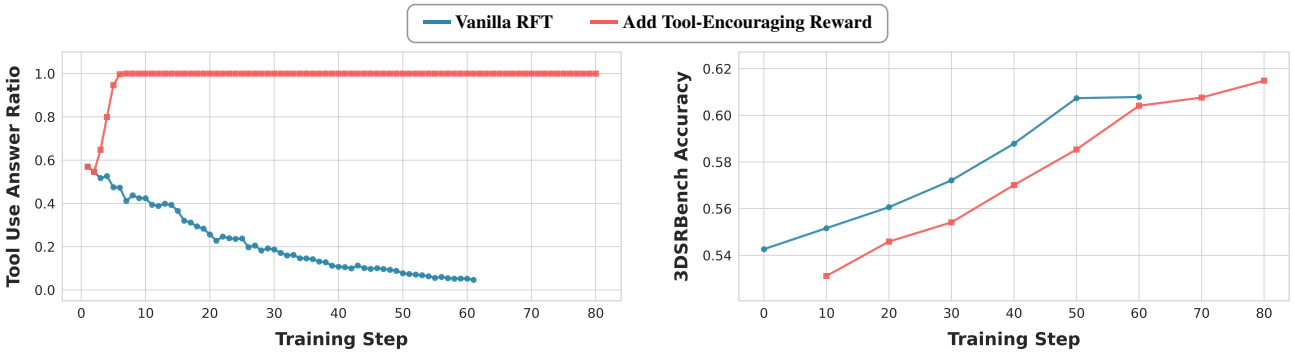
Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Wu, Y., and Wang, M. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.-C., et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025a.

Zhang, X., Yuan, X., Huang, D., You, W., Hu, C., Ruan, J., Chen, K., and Hu, X. Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for llm reinforcement learning. *arXiv preprint arXiv:2510.10959*, 2025b.

Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., and Yu, X. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Zhu, M., Zhong, H., Zhao, C., Du, Z., Huang, Z., Liu, M., Chen, H., Zou, C., Chen, J., Yang, M., et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025.



(a) Tool Use Ratio

(b) Evaluation on SAT validation set

Figure A. Analysis of Training Dynamics on SAT. (a) Tool Use Ratio over training defined as the fraction of rollouts that contain at least one grounding action. (b) Validation accuracy trends. Vanilla training (blue) and tool-encourage reward (red) exhibit divergent tool usage behaviors despite similar limited accuracy gains. This confirms that optimization pathologies persist regardless of dataset size.

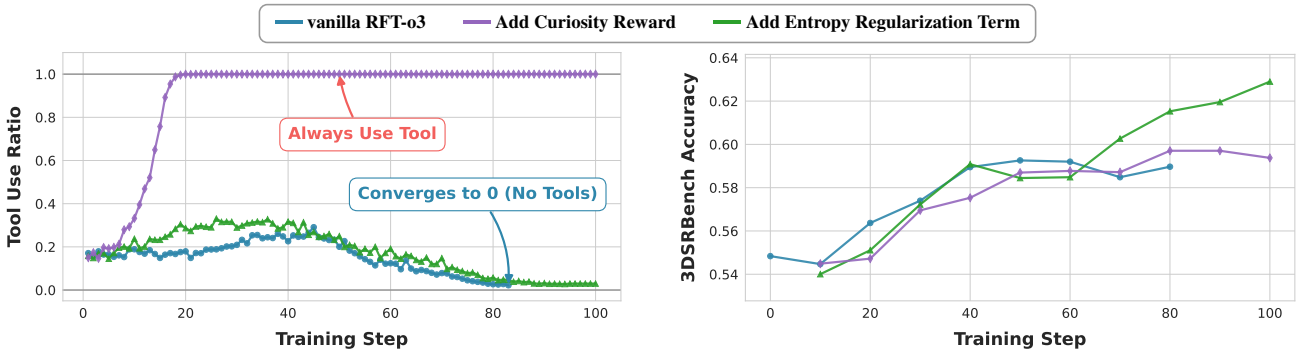
A. Related Works

A.1. Visual Reasoning Agents via Visual Chain-of-Thought

Recent visual agents employ visual tools during chain-of-thought reasoning to enable fine-grained inspection. Most approaches follow a training pipeline of supervised fine-tuning followed by reinforcement learning with GRPO (Su et al., 2025; Wu et al., 2025; Zheng et al., 2025; Wang et al., 2025a; Lai et al., 2025; Zhu et al., 2025; Wang et al., 2025b; Zhang et al., 2025a). These methods can be categorized by their tool repertoire. One category focuses exclusively on zoom-in and cropping operations. DeepEyes (Zheng et al., 2025), Mini-o3 (Lai et al., 2025), Active-o3 (Zhu et al., 2025), Chain-of-Focus (Zhang et al., 2025a), and PixelReasoner (Wang et al., 2025a) train models to adaptively invoke zoom or crop tools through bounding box coordinates, with Mini-o3 employing thought-action-observation loops spanning up to 32 turns. A second category incorporates diverse visual tool sets beyond zooming. OpenThinkIMG (Su et al., 2025) standardizes nine vision tools including object detection, segmentation, and OCR. Simple-o3 (Wang et al., 2025b) demonstrates interleaved vision-language reasoning with cropping, zooming, and image reuse operations. VTool-R1 (Wu et al., 2025) generates multimodal chains of thought using Python-based visual editing tools, while ReFocus (Fu et al., 2025) employs Python code to draw boxes, highlight sections, and mask regions. Visual Sketchpad (Hu et al., 2024) enables test-time tool invocation through sketching and specialist vision models. While these methods achieve strong performance on visual reasoning benchmarks, the training dynamics of tool-use behavior remain underexplored. Our work specifically investigates how RL training shapes tool-use patterns in 3D spatial reasoning tasks where tool utility is ambiguous rather than mandatory, revealing that vanilla RL can drive tool usage toward collapse while exploration-based methods enable tools to function as scaffolding that supports robust tool-free inference.

A.2. Analyzing Visual Chain-of-Thought Behavior

Beyond developing visual CoT methods, recent work has systematically investigated when and how explicit visual information within chain-of-thought reasoning benefits vision-language models. One line of work (Xu et al., 2025) demonstrates that while visual CoT with intermediate edited image patches consistently improves accuracy, it also increases sensitivity to input perturbations, with these intermediate visual reasoning components serving as the primary source of fragility. Another study (Du et al., 2025) investigates how different CoT designs affect generalizable visual reasoning ability, finding that visual CoT with intermediate image generation enables better generalization compared to text-only approaches after reinforcement learning. Our work complements these analyses by investigating the training dynamics of tool-mediated visual reasoning, revealing how exploration diversity determines whether explicit visual tools function as scaffolding or lead to usage collapse.



(a) Tool Use Ratio

(b) Evaluation on 3DSRBench

Figure B. **Analysis of Training Dynamics.** (a) Tool Use Ratio over training, defined as the fraction of rollouts that contain at least one `<grounding>` action. (b) Validation accuracy trends. Vanilla training (blue) and add curiosity reward (purple) exhibit divergent tool usage behaviors despite similar accuracy gains.

B. Additional Experiments

B.1. Training on SAT Dataset

We extend our analysis to the larger SAT dataset (Ray et al., 2024) to verify if the observed tool-use collapse is specific to the smaller 3DSRBench dataset. Figure A presents the training dynamics on SAT. The initial tool-use ratio starts significantly higher at approximately 40% compared to the main experiments. However, the overall trends remain consistent where Vanilla RFT exhibits tool-use collapse as the ratio steadily declines to near zero. Conversely, adding a tool-encouraging reward drives the agent to always use tools. This confirms that these optimization pathologies persist regardless of dataset size or the initial tool-use rate of the policy.

B.2. Training on Add Curiosity Reward

We provide additional experimental results using the Curiosity Reward design (Wang et al., 2025a) discussed in the main text to verify if a more sophisticated reward design mitigates optimization pathology. Figure B illustrates the training dynamics comparing Vanilla RFT, Add Curiosity Reward, and Add Entropy Regularization Term. Consistent with the Tool Bonus results reported in the main text, the Curiosity Reward drives the agent to rapidly saturate tool usage. The tool-use ratio quickly escalates to near 100% within the first 20 training steps. While this approach effectively prevents tool-use collapse, it results in the opposite extreme of indiscriminately invoking tools for every query. Consequently, the validation accuracy shows only marginal improvements and plateaus at a lower level compared to the significant gains achieved by our adaptive entropy regularization method. This reinforces our conclusion that extrinsic reward engineering tends to force rigid tool-use behaviors rather than fostering the diverse exploration required for robust reasoning.

C. Extra Visualization

C.1. Tool-Use and No Tool Example

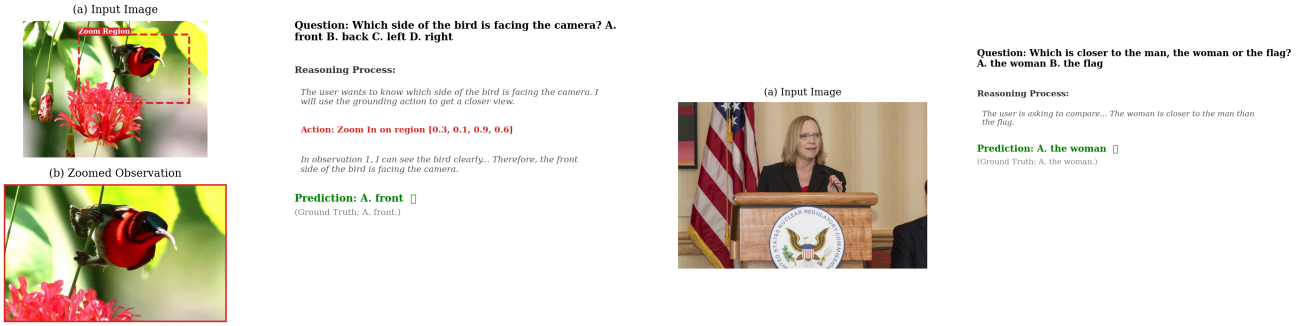
Figure C provides qualitative examples of model rollouts illustrating both multi-turn tool usage and direct tool-free inference.

C.2. Benchmark Difference Between Visual Search and 3D Spatial Understanding Benchmarks

As illustrated in Figure D, visual search benchmarks are heavily skewed toward extremely small objects, implicitly making tool usage mandatory for detection. In contrast, 3D spatial understanding covers a wide spectrum of object scales, which necessitates the agent to learn a selective strategy to determine whether fine-grained inspection is required for relational reasoning or if the global view suffices.

C.3. Token Length Evolution During Training

We analyze the average length of the `<think>` block to assess the depth of reasoning during training. The tool-encouraging reward leads to a progressive decline in token count which suggests the model shortcuts reasoning to greedily pursue tool



(a) Multi-Turn tool usage (Case with zoom-in) *(b)* Direct Answer (Case without Grounding)

Figure C. Comparison of reasoning processes. (a) An example where the model performs a grounding action (zooming in) to identify the bird's orientation. (b) An example where the model answers directly based on the original image.

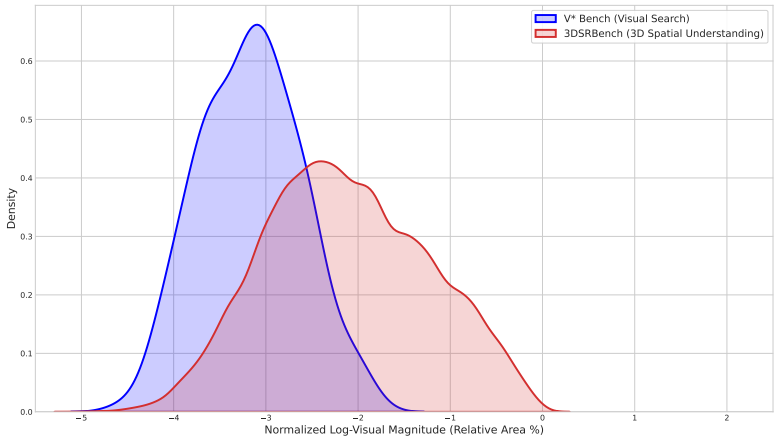
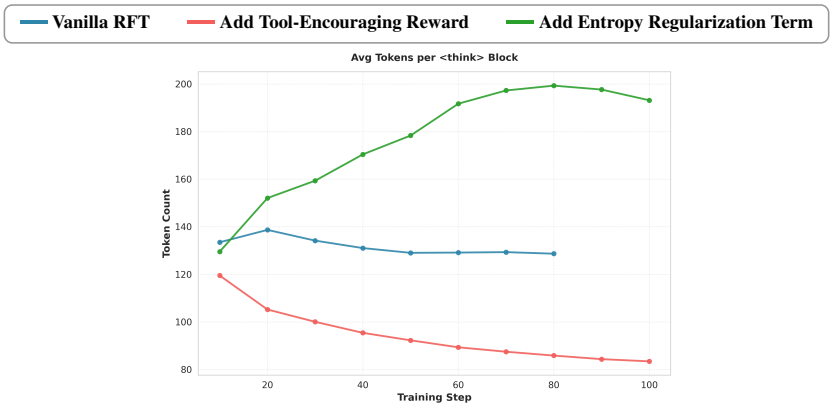


Figure D. Comparison of Target Object Scales. The figure plots the density distribution of normalized log-visual magnitude for target objects in V* (blue) (Wu & Xie, 2024) and 3DSRBench (red) (Ma et al., 2025a). While visual search tasks primarily focus on detecting extremely small objects, 3D spatial reasoning involves a broader range of object sizes. This indicates that the core challenge in 3D spatial reasoning lies not merely in detection but in understanding the relative spatial relationships among objects of various scales.

usage. In contrast, adding the entropy regularization term induces a sustained increase in generation length which indicates that fostering exploration encourages the model to develop richer internal reasoning rather than converging to simplistic behaviors.

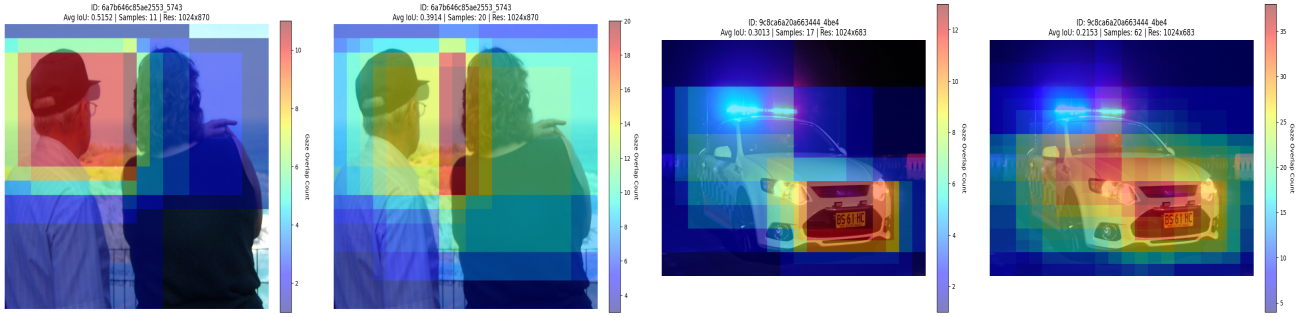
C.4. Additional Visual Exploration Visualization

Figure F presents additional qualitative comparisons of crop heatmaps between the vanilla RFT baseline and the entropy-regularized model during early training. These examples further demonstrate that the entropy-regularized policy maintains broader visual exploration, effectively attending to contextually relevant peripheral regions, whereas the baseline tends to collapse prematurely into a narrow focus on the primary subject.



(a) Avg Tokens per `<think>` Block

Figure E. **Analysis of Training Dynamics.** (a) The average number of tokens within the `<think>` block across training steps. *Vanilla RFT* (blue) exhibits an initial increase followed by stabilization around 130 tokens. The *Tool-Encouraging Reward* (red) leads to a continuous decrease in the token count, suggesting shorter reasoning processes. In contrast, the *Entropy Regularization Term* (green) encourages exploration, resulting in a sustained increase in the length of the `<think>` block.



(a) **Vanilla RFT** *(b)* **Entropy-Reg.** *(c)* **Vanilla RFT** *(d)* **Entropy-Reg.**

Q: *Are the baseball hat and the person facing same or similar directions?*

Q: *Which object is further away?*
A. the police B. the license plate

Figure F. **Additional Visual Exploration During Training Examples.** Comparison of crop distributions. The baseline models (a, c) focus narrowly, while entropy-regularized models (b, d) explore contextually relevant regions.