# DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning

**Ziwei Zheng**[1,2*], **Michael Yang**[1*], **Jack Hong**[1*], **Chenxiao Zhao**[1*†], **Guohai Xu**[1‡], **Le Yang**[2‡], **Chao Shen**[2], **XingYu**[1]
[1]Xiaohongshu Inc., [2] Xi'an Jiaotong University

Project Homepage

{chenxiao2, xuguohai}@xiaohongshu.com, ziwei.zheng@stu.xjtu.edu.cn,
{yangminghao199,jaaackhong}@gmail.com, yangle15@xjtu.edu.cn

## Abstract

Large Vision-Language Models (VLMs) have shown strong capabilities in multimodal understanding and reasoning, yet they are primarily constrained by text-based reasoning processes. However, achieving seamless integration of visual and textual reasoning which mirrors human cognitive processes remains a significant challenge. In particular, effectively incorporating advanced visual input processing into reasoning mechanisms is still an open question. Thus, in this paper, we explore the interleaved multimodal reasoning paradigm and introduce ***DeepEyes***, a model with "thinking with images" capabilities incentivized through end-to-end reinforcement learning without the need for cold-start SFT. Notably, this ability emerges natively within the model itself, leveraging its inherent grounding ability as a tool instead of depending on separate specialized models. Specifically, we propose a tool-use-oriented data selection mechanism and a reward strategy to encourage successful tool-assisted reasoning trajectories. *DeepEyes* achieves significant performance gains on fine-grained perception and reasoning benchmarks and also demonstrates improvement in grounding, hallucination, and mathematical reasoning tasks. Interestingly, we observe the distinct evolution of tool-calling behavior from initial exploration to efficient and accurate exploitation, and diverse thinking patterns that closely mirror human visual reasoning processes. Code is available at `https://github.com/Visual-Agent/DeepEyes`.

## 1 Introduction

Recent advances in Vision-Language Models (VLMs) have enabled deeper reasoning over multimodal inputs by adopting long Chain-of-Thought (CoT) approaches [1, 2, 3], allowing these models to handle more complex tasks. However, these models still primarily rely on text-based reasoning, with their thought processes largely confined to the language modality. In contrast, human reasoning naturally combines vision and cognition, thinking with images by extracting information through sequential visual fixations, which support more accurate perceptual decision-making, being essential for survival in early human evolution [4]. While some recent works have proposed pre-defined workflow-based strategies to incorporate visual information into CoT reasoning [5, 6], the modular designs suffer from suboptimal performance [7].

*Equal contribution. Listing order is random. Work done during Ziwei's internship at Xiaohongshu.
†Main Code Contributor
‡Corresponding Author

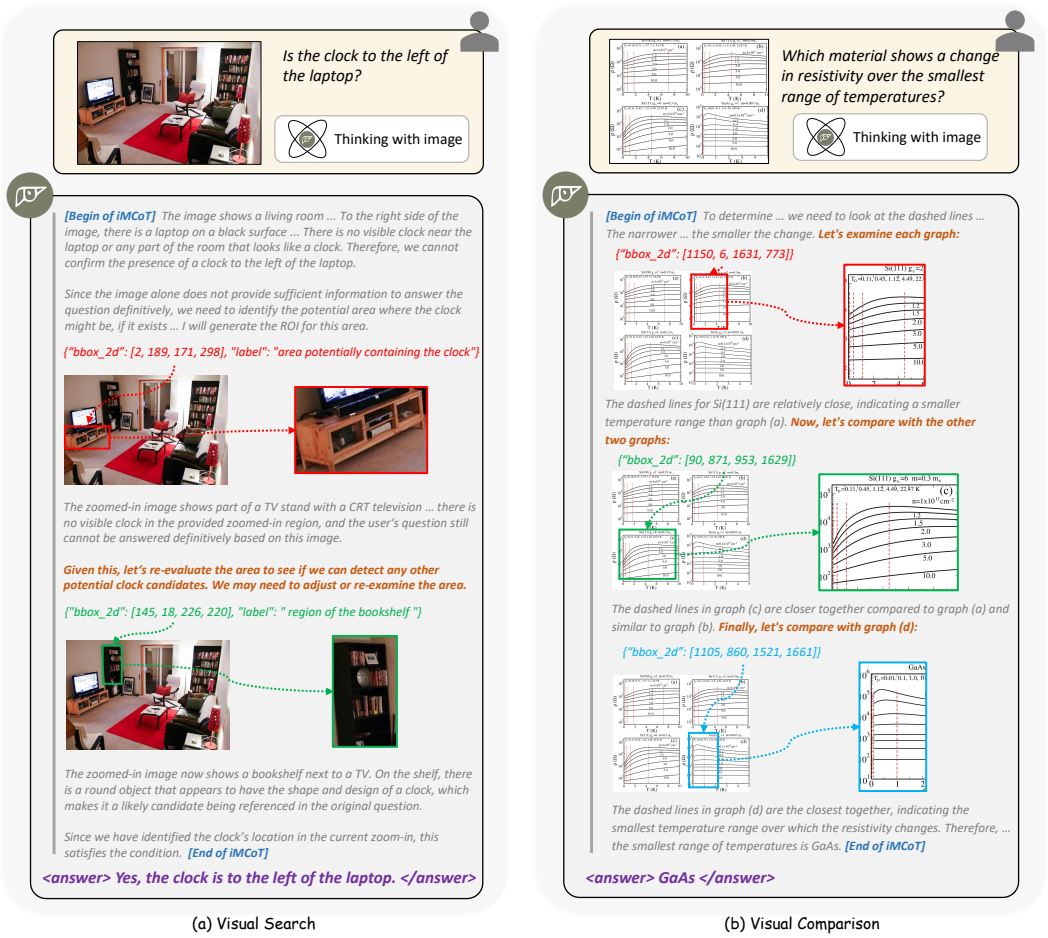



Figure 1: **Interleaved Multi-modal Chain-of-Thought (iMCoT).** *DeepEyes* is incentivized to perform active perception throughout the reasoning process with end-to-end reinforcement learning.

In the recent milestone development, the OpenAI o3 model [8], visual information has been successfully integrated as a dynamic element in the reasoning process. The o3 transcends the language-modality confinement by extending reasoning capability to "thinking with images" like humans. Additionally, it resolves the coordination limitations by combining textual CoT and image manipulation tools in a naturally interleaved fashion during the CoT process. This approach enables a new axis for test-time compute scaling by seamlessly integrating visual and textual reasoning, representing a meaningful advancement toward true multimodal reasoning. However, the inner mechanism still remains undisclosed to the open-source community.

In this paper, we introduce ***DeepEyes***, a model with "thinking with images" ability, which is incentivized through end-to-end reinforcement learning. This capability emerges natively without relying on separate specialized models. It is directly guided by outcome reward signals, thereby eliminating the need for cold-start supervised fine-tuning by previous methods. Specifically, we encapsulate the model's grounding ability in an image zoom-in tool, enabling it to actively gather information from the original image by calling tool functions in an agentic framework. As shown in Figure 1, the model adaptively generates image grounding coordinates and crops relevant regions, which are then concatenated into the ongoing reasoning trajectory. This allows an interleaved Multimodal Chain-of-Thought (iMCoT), where visual and textual reasoning are integrated seamlessly.

In early attempts, we observe that the model is initially reluctant to use the image zoom-in tool. Besides, early exploration often results in poorly chosen zoom-in regions and low rewards, leading to instability in training dynamics. To address these issues, we propose a data selection mechanism to choose training samples based on their potential to encourage tool-calling behavior. Additionally, we design a reward strategy that assigns a conditional tool-usage bonus to the trajectories that successfully

complete their tasks through tool calling. Our ablation and analysis show that the above two strategies help ensure the efficiency and accuracy of tool usage are properly optimized.

Whereas we do not apply any supervised fine-tuning for intermediate steps, we observe the tool-calling dynamics evolve through three distinct stages during RL training: (1) initial exploration with limited effectiveness, followed by (2) aggressive yet successful usage, and finally (3) developing selective and efficient exploitation with high performance. The tendency demonstrates progressive mastery of visual reasoning capabilities with tool usage. Additionally, diverse iMCoT reasoning patterns emerge, such as visual search for small or hard-to-recognize objects, visual comparisons across different regions, visual confirmation to eliminate uncertainty, and hallucination mitigation through focusing on details. These diverse reasoning behaviors closely resemble human cognitive processes, thereby enhancing the system's overall multimodal capabilities.

Experimental results show that *DeepEyes* can significantly boost performance on multiple visual perception and reasoning tasks. For high-resolution benchmarks, *DeepEyes* with a 7B model achieves an accuracy of 90.1% (+18.9 %) on $V^*$, and improves HR-Bench-4K and HR-Bench-8K by 6.3% and 7.3%, respectively. In addition, *DeepEyes* also improves multimodal capabilities on a wide range of tasks such as visual grounding, hallucination mitigation, and math problem solving.

The main contributions are summarized as follows:

- We incentivize and enhance the ability of thinking with images via end-to-end reinforcement learning, forming iMCoT that seamlessly blends visual-textual reasoning without requiring cold-start SFT or separate specialized models as external tools.
- To better incentivize the model's reasoning behavior, we introduce a combination of tool-use-oriented data selection and a reward strategy that strongly encourages tool-assisted problem solving. Experiments show that both components significantly contribute to the advancement of iMCoT.
- We reveal the intriguing RL training dynamic of iMCoT, where tool-calling behavior undergoes distinct stages, evolving from initial tool exploration to efficient and accurate tool exploitation. We also observe diverse reasoning patterns, such as visual search, comparison, and confirmation.

## 2 Related Work

**Multi-modal Large Language Models.** Multimodal large language models (MLLMs) have evolved from early systems that loosely combined vision encoders with language models into more integrated architectures through joint training. Methods such as BLIP-2 [9] and LLaVA [10, 11] align visual and linguistic modalities by projecting image features into the latent space of frozen LLMs using query transformers or lightweight projectors, enabling tasks like visual question answering and instruction following. To address resolution constraints, approaches like AnyRes [12, 13] allow for flexible image sizes and enhanced visual fidelity. These advances have led to strong open-source models, including the LLaVA [14, 15, 16, 17, 18], Qwen-VL [19, 20, 21], and InternVL [22, 23, 24] series. Concurrently, large-scale models like Flamingo [25], mPLUG-Owl [26, 27, 28], and GPT-4V [29] aim to unify vision-language understanding, incorporating mechanisms such as mixture-of-experts [30, 31, 32] or image generation [33, 34]. However, these models lack reasoning capabilities like Chain-of-Thought and test-time scalability [35, 36, 37], and still decouple perception from reasoning.

**Vision-language Model Reasoning.** Existing Multimodal Chain-of-Thought (MCoT) reasoning methods fall into two main categories. Early approaches rely on predefined workflows, staged procedures, or auxiliary models [38, 39, 40], often focusing on region-of-interest localization [41, 42, 43, 44], latent feature regeneration [45, 46], and external knowledge integration [6, 47] to improve interoperability. Inspired by the extensive research on the long CoT in LLMs [48], RL-based reasoning approaches have been increasingly explored in MLLMs [49, 50, 51]. These methods predominantly extend text-only reasoning capabilities to a range of multimodal tasks such as spatial reasoning [52], object recognition [53], and semantic segmentation [54]. Unlike prior methods that either hard-code reasoning pipelines or directly extend text-only CoT, our approach enables the model to autonomously decide when and how to incorporate visual input. Guided by outcome reward signals, it adaptively adjusts visual exploration during reasoning, yielding a more flexible process.

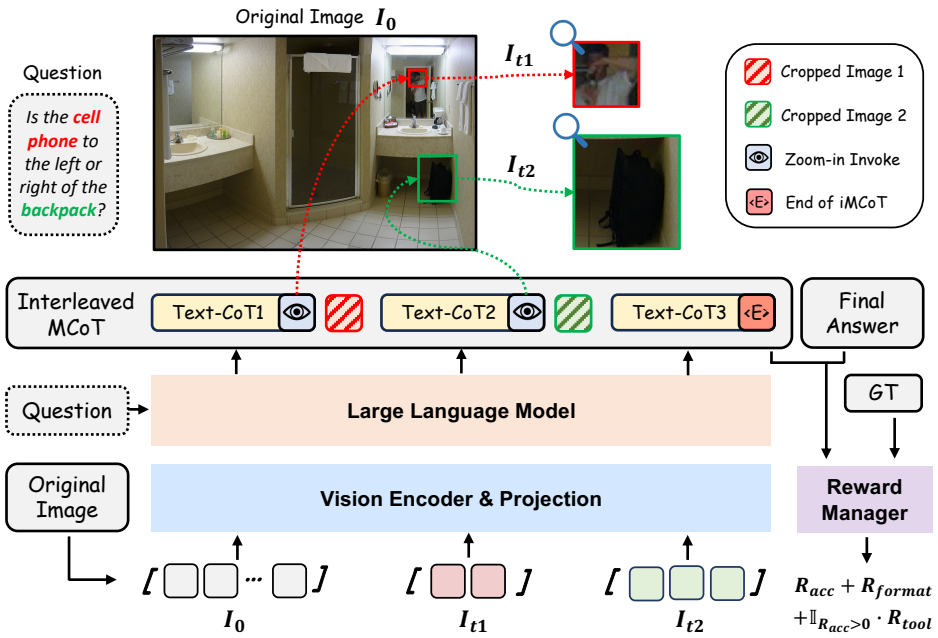



Figure 2: **Overview of *DeepEyes*.** Our model itself decides whether to perform a second perception via zoom-in by generating grounding coordinates and cropping relevant regions, or to answer directly.

# 3 Method

In this section, we first present an overview of the proposed *DeepEyes* in Section 3.1. The end-to-end RL procedure along with the corresponding reward design are introduced in 3.2. Section 3.3 introduces our data collection and data selection mechanism.

## 3.1 DeepEyes

*DeepEyes* is a unified multimodal agent that is capable of "thinking with images" through an iMCoT reasoning process. The ability is inherited from the model's native capability of visual grounding and action decision planning, and further incentivized and enhanced via end-to-end RL training using outcome reward signals, eliminating the need for cold-start supervised fine-tuning.

As illustrated in Figure 2, given a user question and an image $I_0$ as input, *DeepEyes* can autonomously decide, after each textual CoT reasoning step, whether to directly draw an answer, or invoke a image zoom-in function for further image inspection. The image zoom-in function takes a list of bounding box coordinates as input, and outputs the cropped images within the specified regions. The returned cropped images, such as $I_{t1}$ and $I_{t2}$, are appended to the ongoing trajectory, allowing the model to reason over all previous context. *DeepEyes* can invoke the image zoom-in function as many times as needed before concluding a final answer. This iterative interaction enables fine-grained perception, especially when the relevant object in the image is small, blurry, or difficult to recognize. During the RL training stage, the reward optimization policy gradient is applied to the entire trajectory, allowing all textual CoTs and action decision planning can be jointly optimized in an end-to-end manner.

Compared to previous works based on workflows or pure text reasoning, our iMCoT offers several significant advantages. (1) **Simplicity in Training.** Previous workflow-based methods depend on substantial SFT data, which is challenging to acquire, while our iMCoT only requires question-answer pairs, reducing data collection complexity. (2) **Enhanced Generalizability.** Workflow-based models are constrained by their task-specific manual design constraints, leading to generalization to other tasks. In contrast, our iMCoT exhibits robust generalization capabilities as it learns to dynamically select optimal reasoning processes across diverse tasks through reinforcement learning. (3) **Unified Optimization.** Our iMCoT enables joint optimization through end-to-end training, which ensures global optimality of the system. In contrast, optimizing each component separately typically leads to sub-optimal performance. (4) **Multimodal Integration.** Compared to pure text-based thinking, our iMCoT naturally interleaves visual and textual information, combining visual elements with

textual reasoning to achieve more accurate perceptual decision-making. (5) **Native Tool Calling.** The native "thinking with images" ability allows for direct optimization of tool utilization efficiency and accuracy, which can not be achieved by previous reasoning paradigms.

### 3.2 Agentic Reinforcement Learning

**Rollout Formulation.** In traditional RL with text-only CoT, the Markov Decision Process (MDP) defines the state as the input prompt tokens together with all tokens generated by the model up to the current step. The action is defined as the next token in the sequence. In contrast, agentic RL extends this formulation by introducing observation tokens, which come from external function calls rather than the model itself. These observation tokens are appended to the ongoing rollout sequence and fed back into the model as input for the subsequent step.

We formalize the MDP definition for iMCoT as follows. At each step $t$, the state $s_t$ of iMCoT is defined as:

$$s_t = \{(X_0, I_0), (X_1, I_1), \ldots, (X_t, I_t)\} = \{\mathbf{X}_{\leq t}; \mathbf{I}_{\leq t}\}, \tag{1}$$

where $\mathbf{X}_{\leq t} = \{X_1, \ldots, X_t\}$ represents the accumulated sequence of text tokens before step $t$, and $\mathbf{I}_{\leq t} = \{I_1, \ldots, I_t\}$ represents the image observation tokens before step $t$. We omit other related special tokens that are not generated by VLM itself for simplicity. Given the state $s_t$, the action $a_t \sim \pi_\theta(a \mid s_t)$ is sampled from the VLM policy $\pi_\theta$, serving as the next input token. This iMCoT continues to interleave until either an answer is generated or the maximum number of tool calls is reached. Note that the text tokens $\mathbf{X}_{\leq \mathbf{t}}$ and image tokens $\mathbf{I}_{\leq \mathbf{t}}$ are interleaved in the states. All observation tokens are considered as a whole, which does not contribute to the loss computation.

**Reward Design.** In multimodal environments, sparse and outcome-driven reward signals are crucial for guiding vision-language models toward effective reasoning and decision-making. Given the lack of step-level supervision in intermediate visual actions, we adopt a reward formulation that evaluates the reasoning trajectory based on final outcome quality and strategic tool usage.

The total reward is composed of three parts: an accuracy reward $R_{\text{acc}}$, a formatting reward $R_{\text{format}}$, and a conditional tool usage bonus $R_{\text{tool}}$. The accuracy reward assesses whether the final answer is correct, while the formatting reward penalizes poorly structured outputs. The tool usage bonus is awarded only when the model produces a correct answer and invokes at least one external perception tool during the trajectory. Formally, given a reasoning trajectory $\tau$, the total reward is defined as:

$$R(\tau) = R_{\text{acc}}(\tau) + R_{\text{format}}(\tau) + \mathbb{I}_{R_{\text{acc}}(\tau)>0} \cdot R_{\text{tool}}(\tau), \tag{2}$$

where $\mathbb{I}_{R_{\text{acc}}(\tau)>0}$ is the indicator function which takes the value of 1 only when $R_{\text{acc}}(\tau) > 0$. We find that directly rewarding the model for tool usage is essential to promote perception-aware reasoning (see Section 4.3). The tool reward is *conditional*: it only applies when the final answer is correct and at least one tool is used during the trajectory. This encourages the model to invoke tools meaningfully when they contribute to successful task completion, rather than as arbitrary or redundant actions.

**Optimization.** For the RL algorithm, we adopt Group Relative Policy Optimization (GRPO) [55], which has been proved to be effective and efficient for diverse tasks. For multi-turn agent trajectories, we apply a token-wise loss mask to ignore loss on observation tokens not generated by the model.

### 3.3 Training Data

**Data Collection.** Our data collection adheres to three fundamental principles: (1) Diverse Tasks and Image Distribution. We incorporate varied data to enhance the generalization capabilities of our iMCoT. (2) Tool Effectiveness. We select scenarios where tool usage demonstrably improves accuracy. (3) Reasoning Ability Enhancement. We carefully choose data that effectively improves the model's reasoning capabilities. Consequently, our training dataset comprises three complementary components: fine-grained data, chart data, and reason data. Fine-grained data is selected from part of $V^*$ training set [41], focusing on high-resolution images and detailed perception questions that maximize tool effectiveness. Chart data from ArxivQA [56] features synthetic charts and graph images that enhance the diversity of visual elements. For the reasoning data, we incorporate the ThinkLite-VL [57] dataset to broaden task diversity and strengthen the model's analytical capabilities. More detailed analysis of our data is provided in Appendix B.

Table 1: **Results on High-Resolution Benchmarks.** E2E indicates whether the model is end-to-end, requiring no manually defined workflow. * denotes the results are reproduced by ourselves.

| Model | E2E | Param Size | $V^*$ Bench [41] | | | HR-Bench 4K [59] | | | HR-Bench 8K [59] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Attr | Spatial | Overall | FSP | FCP | Overall | FSP | FCP | Overall |
| GPT-4o [60] | ✓ | - | - | - | 66.0 | 70.0 | 48.0 | 59.0 | 62.0 | 49.0 | 55.5 |
| o3 [8] | ✓ | - | - | - | 95.7 | - | - | - | - | - | - |
| SEAL [41] | ✗ | 7B | 74.8 | 76.3 | 75.4 | - | - | - | - | - | - |
| DyFo [44] | ✗ | 7B | 80.0 | 82.9 | 81.2 | - | - | - | - | - | - |
| ZoomEye [61] | ✗ | 7B | 93.9 | 85.5 | 90.6 | 84.3 | 55.0 | 69.6 | 88.5 | 50.0 | 69.3 |
| LLaVA-OneVision [62] | ✓ | 7B | 75.7 | 75.0 | 75.4 | 72.0 | 54.0 | 63.0 | 67.3 | 52.3 | 59.8 |
| Qwen2.5-VL* [58] | ✓ | 7B | 73.9 | 67.1 | 71.2 | 85.2 | 52.2 | 68.8 | 78.8 | 51.8 | 65.3 |
| Qwen2.5-VL* [58] | ✓ | 32B | 87.8 | 88.1 | 87.9 | 89.8 | 58.0 | 73.9 | 84.5 | 56.3 | 70.4 |
| **DeepEyes** | ✓ | 7B | 91.3 | 88.2 | 90.1 | 91.3 | 59.0 | 75.1 | 86.8 | 58.5 | 72.6 |
| Δ (*vs* Qwen2.5-VL 7B) | - | - | +17.4 | +21.1 | +18.9 | +6.1 | +6.8 | +6.3 | +10.0 | +6.8 | +7.3 |

Table 2: **Results on Grounding and Hallucination Benchmarks.** We compare *DeepEyes* with open-source MLLMs on several grounding and hallucination benchmarks. * denotes the results are reproduced by ourselves.

| Model | Param Size | refCOCO | refCOCO+ | refCOCOg | ReasonSeg | POPE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Adversarial | Popular | Random | Overall |
| LLaVA-OneVision [62] | 7B | - | - | - | - | - | - | - | 88.4 |
| Qwen2.5-VL [58] | 7B | 90.0 | 84.2 | 87.2 | - | - | - | - | - |
| Qwen2.5-VL* [58] | 7B | 89.1 | 82.6 | 86.1 | 68.3 | 85.9 | 86.5 | 87.2 | 85.9 |
| **DeepEyes** | 7B | 89.8 | 83.6 | 86.7 | 68.6 | 84.0 | 87.5 | 91.8 | 87.7 |
| Δ (*vs* Qwen2.5-VL 7B) | - | +0.7 | +1.0 | +0.6 | +0.3 | -1.9 | +1.0 | +4.6 | +1.8 |

Table 3: **Results on Multimodal Reasoning Benchmarks.** We evaluate our model on several multimodal reasoning benchmarks. * denotes the results are reproduced by ourselves, and † represents the results are copied from [63].

| Model | Param Size | Math Vista [64] | Math Verse [65] | Math Vision [66] | We Math [67] | Dyna Math [68] | Logic Vista [69] |
|---|---|---|---|---|---|---|---|
| LLaVA-OneVision [62] | 7B | 58.6† | 19.3† | 18.3† | 20.9† | - | 33.3† |
| Qwen2.5-VL [58] | 7B | 68.2 | 49.2 | 25.1 | 35.2† | - | 44.1† |
| Qwen2.5-VL* [58] | 7B | 68.3 | 45.6 | 25.6 | 34.6 | 53.3 | 45.9 |
| **DeepEyes** | 7B | 70.1 | 47.3 | 26.6 | 38.9 | 55.0 | 47.7 |
| Δ (*vs* Qwen2.5-VL 7B) | - | +1.9 | +1.7 | +1.0 | +4.3 | +1.7 | +1.8 |

**Data Selection.** Our tool-use-oriented data selection strategy encompasses four key steps: (1) Managing Difficulties: we use Qwen2.5-VL-7B [58] to generate 8 responses per question and estimate the difficulty based on the accuracy. Samples with accuracy of either 0 or 1 are excluded as they are either too challenging or too elementary for effective learning. (2) Structuring Question Formats: We reformulate original questions into the open-ended format, excluding those that cannot be reliably converted. (3) Ensuring Verifiability: We eliminate data that cannot be properly verified, such as questions with incorrect answers or those that are unreadable. (4) Facilitating Tool Integration: We implement an additional filtering step to prioritize samples offering higher information gain through tool-calling. We select instances where the model produces incorrect answers in single-turn interactions but achieves correct results when utilizing ground-truth crop regions, highlighting cases where visual tool use is most beneficial. Specifically, chart data is exempted from the tool integration filtering process, while reason data retains its original form as it has already undergone rigorous processing. Through this comprehensive selection strategy, we curate a high-quality dataset specifically optimized for developing and enhancing tool-aware visual reasoning capabilities.

# 4 Experiment

## 4.1 Setups

**Baselines and Benchmarks.** To comprehensively assess the effectiveness of *DeepEyes*, we compare it against three categories of baselines: (1) advanced *proprietary* models, including GPT-4o [60] and o3 [8]; (2) state-of-the-art *open-source* models, such as LLaVA-OneVision [62] and Qwen2.5-VL [58]; and (3) approaches explicitly designed with *workflows*, such as SEAL [41], DyFo [44] and ZoomEye [61]. Since tasks requiring fine-grained visual understanding naturally highlight the

strengths of iMCoT, we first evaluate *DeepEyes* on high-resolution benchmarks. Then, we assess *DeepEyes* on grounding and hallucination benchmarks to show improvements brought by iMCoT on general visual capabilities. We also adopt general reasoning benchmarks to verify its effectiveness.

**Training Details.** We train Qwen2.5-VL-7B with GRPO for 80 iterations on H100 GPUs. Each batch samples 256 prompts, with 16 rollouts per prompt, up to a maximum of 6 times of tool calling. We set the KL coefficient to 0.0 and define the maximum response length as 20480 tokens.

### 4.2 Main Results

**High-Resolution Benchmarks.** High-resolution benchmarks, including $V^*$ [41] and HR-Bench [59], feature images with very high resolutions ranging from 2K to 8K. Additionally, the target objects referred to in the questions are often quite small in these images, potentially occupying only one or two hundred pixels. The extreme resolution combined with the small size of objects makes it extremely challenging for VLMs to accurately locate the target objects, frequently resulting in incorrect responses. As shown in Table 1, our model achieves exceptional performance on high-resolution benchmarks and significantly outperforms existing open-source models, with results that surpass even manually engineered and complex workflows [41, 44, 61]. Compared to Qwen2.5-VL 7B, our model achieves remarkable performance gains of $18.9\%$ and $7.3\%$ on $V^*$ Bench [41] and HR-Bench 8K [59] respectively, which demonstrates the critical importance of visual reasoning ability of "thinking with images" for high-resolution perception. Notably, without complex pipeline design or an elaborate training process, we can successfully unlock this capability through simple RL.

**Grounding and Hallucination Benchmarks.** Furthermore, the multimodal CoT can also enhance general visual capabilities. We evaluate our model on grounding (refCOCO [70], refCOCO+ [70], refCOCOg [71], and ReasonSeg [72]), and hallucination (POPE [73]) benchmarks. From Table 2, our model achieves higher accuracy on the grounding task and shows substantial improvement in reducing hallucinations. This improvement stems from our model's ability to focus on specific regions of interest during the visual reasoning process and perform detailed analysis of these cropped areas, thereby more confidently confirming the presence or absence of objects. The results highlight that iMCoT not only improves high-resolution perception but also enhances the model's general reliability by providing a more thorough focus and verification mechanism for visual content.

**Multimodal Reasoning Benchmarks.** Moreover, our model also demonstrates strong reasoning ability. We evaluate our model on several reasoning benchmarks and compare the results with previous models in Table 3. Due to the introduction of chain-of-thought, our model achieves generalized performance improvements across several multimodal reasoning benchmarks.

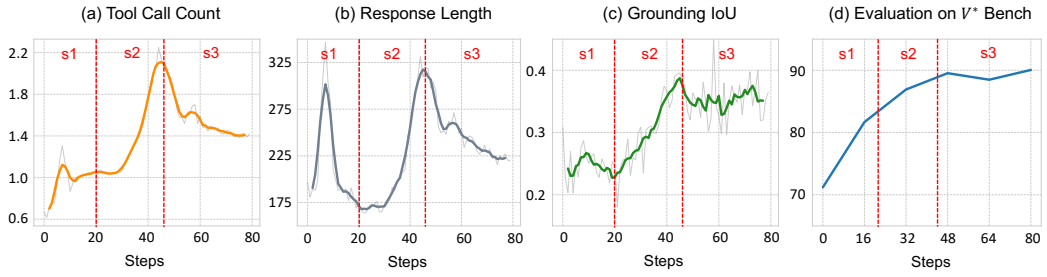



Figure 3: **Training dynamics of *DeepEyes*.** s1/2/3 represent different stages.

### 4.3 Key Findings: From Casual Tool User to Proficient Tool Master

**Training Dynamics.** To gain deeper insights into the model's behavior throughout end-to-end reinforcement learning, we perform a detailed analysis on fine-grained data. Given that fine-grained data includes ground-truth bounding boxes closely aligned with the target answers, we quantify the quality of the accuracy of model cropping using Intersection-over-Union (IoU). In Figure 3, we observe a clear shift in how the model interacts with tools. This evolution unfolds in three distinct stages, each reflecting a progressively more effective integration of tool use into model's reasoning:

• **Stage 1: Initial Tool Exploration (Steps 0–20)** The model begins by responding to system prompts to invoke tools, but lacks a coherent or effective usage policy. Both tool call count and response

length increase noticeably, signaling exploratory behavior. However, grounding IoU remains low, indicating that the model often invokes tools without successfully linking retrieved information to visual context. This stage is marked by reactive trial-and-error behavior, where the model is primarily probing the utility of available tools without external guidance. Interestingly, a sharp drop in response length occurs between steps 8 and 20, as the model begins to trim down verbose image descriptions and tool-related intent statements while acquiring basic tool-use skills.

• **Stage 2: High-Frequency Tool Usage (Steps 20–45)** The model enters a phase of aggressive tool usage, repeatedly invoking tools to maximize both answer correctness and tool rewards. This strategy yields substantial gains across all key performance metrics, including grounding IoU and accuracy. The longer responses and higher tool call frequency suggest a "broad sweep" strategy: rather than relying on internal reasoning, the model externalizes visual reasoning by over-querying the environment. This reflects a transitional stage where the model begins to recognize the functional relevance of tools but has not yet learned to use them efficiently.

• **Stage 3: Efficient Tool Exploitation (Steps 45–80)** The model shifts toward more selective and precise tool use, reducing both tool invocation frequency and response length, while maintaining high grounding and task accuracy. This behavior indicates that the model has internalized a more compact visual-linguistic policy: it now uses tools not as a "crutch" but as a complementary resource, invoked only when necessary. The high grounding IoU with fewer tool calls reflects the emergence of an implicit planning mechanism, wherein the model first narrows down the likely visual scope internally before selectively leveraging tools to confirm or refine its hypothesis.

The training evolves from broad exploration to targeted exploitation, demonstrating the model's ability to learn tool use and optimize it for reward. Tool use becomes an integral part of the model's reasoning, co-evolving with its policy through end-to-end training. These strategies underscore the potential of tool-augmented visual-language models for scalable, interpretable multimodal reasoning.

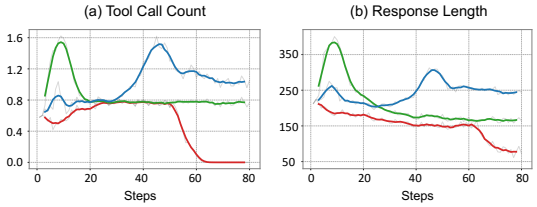



Figure 4: Training dynamics w.r.t. tool reward.

Figure 5: Evaluations w.r.t. tool reward.

| Method | $V^*$ | HR-4k | HR-8k |
|---|---|---|---|
| w/o Tool Reward | 87.4 | 53.4 | 55.4 |
| Unconditional Reward | 87.4 | 72.1 | 71.8 |
| Conditional Reward | 90.1 | 75.1 | 72.6 |

**Tool Reward.** The reward design in Eq. 2 includes a conditional tool reward, granting a bonus only when the model correctly answers a question with grounding tools. This mechanism is crucial to encourage effective and purposeful tool use. For comparison, we experiment with removing the tool reward (*w/o tool reward*) and removing its dependency on task accuracy (*unconditional reward*). Training dynamics and evaluation results for these variants are shown in Figure 4 and Table 5. From the results, we observe that when no tool reward is provided, the model quickly reduces its reliance on tools and eventually stops using them altogether. In comparison, introducing a tool reward without conditioning on correctness helps maintain a basic level of tool usage, but the behavior remains static and does not improve over time. The model shows limited motivation to adapt or explore more sophisticated reasoning strategies. On the other hand, when the tool reward is conditioned on producing correct answers, the model progressively increases its tool usage and generates longer, more informative responses, indicating a deeper integration of tools into its reasoning process. These differences in training behavior are clearly reflected in the evaluation results. The conditional tool reward achieves the highest accuracy, outperforming the settings without reward or with unconditional reward. This demonstrates that rewarding tool usage alone is not sufficient; it is the alignment of rewards with meaningful outcomes that truly drives intelligent and effective behavior in *DeepEyes*.

**Thinking Patterns.** Here, we analyze the diverse thinking patterns that emerged during end-to-end RL training, showing how the model integrates visual tools into its reasoning in ways that mirror human visual cognition and adapts their use to task demands. Four primary patterns can be identified:

• **Visual Search** When facing complex problems that a single observation can't solve, the model uses a zoom-in tool to scan different image regions, gathers visual clues, and reasons through them to reach reliable conclusions (Figure 7).

Table 4: **Ablation Study on iMCoT.** We compare the results of RL training using text-only CoT and iMCoT on the same datasets.

| Model | $V^*$ Bench | | | HR-Bench 4K | | | HR-Bench 8K | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Attr** | **Spatial** | **Overall** | **FSP** | **FCP** | **Overall** | **FSP** | **FCP** | **Overall** |
| Qwen2.5-VL [58] | 73.9 | 67.1 | 71.2 | 85.2 | 52.2 | 68.8 | 78.8 | 51.8 | 65.3 |
| RL w. Text-only CoT | 90.4 | 85.5 | 88.5 | 92.3 | 58.5 | 75.4 | 69.3 | 52.3 | 60.8 |
| **DeepEyes** | 91.3 | 88.2 | 90.1 | 91.3 | 59.0 | 75.1 | 86.8 | 58.5 | 72.6 |

Table 5: **Impact of Training Data.** Fine represents the fine-grained data. HR denotes HR-Bench. Row #0 is the origin score of Qwen2.5-VL 7B.

| # | Fine | Reason | Chart | High-Resolution | | | Basic VL Capability | | Reasoning | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $V^*$ Bench | HR-4K | HR-8K | ReasonSeg | POPE | MathVista | MathVerse |
| **0** | | | | 71.2 | 68.8 | 65.3 | 68.3 | 85.9 | 68.2 | 45.6 |
| **1** | ✓ | | | 86.9 | 68.9 | 67.3 | 69.0 | 86.6 | 67.0 | 42.9 |
| **2** | ✓ | | | 91.6 | 74.1 | 71.0 | 69.1 | 88.1 | 64.7 | 41.3 |
| **3** | ✓ | ✓ | | 91.6 | 73.8 | 70.5 | 68.6 | 88.8 | 67.7 | 43.8 |
| **4** | ✓ | | ✓ | 90.1 | 74.6 | 74.6 | 68.5 | 87.9 | 64.6 | 38.1 |
| **5** | ✓ | ✓ | ✓ | 90.1 | 75.1 | 72.6 | 68.6 | 87.7 | 70.1 | 47.3 |

• **Visual Comparison** When handling fine-grained understanding across multiple images or objects, the model iteratively zooms in on each one, allowing close examination and comparison before drawing a final conclusion (Figure 8).

• **Visual Confirmation** In some cases, the model begins with uncertainty but gradually builds confidence by zooming in on image details to gather evidence and resolve doubts (Figure 9).

• **Hallucination Mitigation** Although vision-language models can sometimes hallucinate, invoking the zoom-in tool helps the model focus on visual details to mitigate hallucination. (Figure 10).

### 4.4 Ablation Study

**Interleaved Multimodal Chain-of-Thought.** To validate the effectiveness of iMCoT, we train Qwen2.5-VL-7B based on text-only CoT via end-to-end RL using the same training datasets. As shown in Table 4, text-only CoT reasoning also improves perception ability. However, we observe inferior performance on extremely high-resolution images (HR-Bench 8K). This limitation may arise from the absence of such image dimensions in the training data, causing these images to be treated as out-of-distribution samples and thereby reducing the model's effectiveness. In contrast, *DeepEyes* seamlessly integrates tool utilization into the reasoning process, effectively overcoming resolution constraints and demonstrating superior performance in analyzing extremely high-resolution images.

**Training Data.** We further investigate the influence of different training data and report results in Table 5. Firstly, training with fine-grained data (#2) significantly enhances the model's ability to handle high-resolution images, while experiments on unfiltered data (#1) yield minimal gains (0.1% on HRBench-4K), underscoring the necessity and effectiveness of our tool-use-oriented data selection. Besides, training exclusively on perception data leads to catastrophic forgetting of reasoning abilities. The addition of reasoning data (#3) partially mitigates this forgetting problem, helping the model retain some of its mathematical reasoning capabilities while still improving performance on high-resolution perception tasks. Furthermore, the inclusion of chart data (#4) expands the diversity of training images, introducing new visual elements and structures. Many questions in chart data involve understanding relationships between multiple image elements, which extends the variety of tasks. This increased diversity further enhances the model's thinking capabilities by exposing it to more complex relational reasoning scenarios. The results in row (#2, 3, 4) clearly illustrate the complementary benefits of each data types. While high-resolution data improves the perception ability, reasoning data helps preserve critical reasoning capabilities, and chart data broadens the understanding of visual relationships. Therefore, we combine these complementary data sources (#5) to more effectively and comprehensively activate the model's visual thinking capabilities.

## 5 Conclusion

In this paper, we introduce *DeepEyes*, a vision-language model trained via end-to-end reinforcement learning to blend visual inputs with textual reasoning seamlessly, forming iMCoT. Unlike prior methods, *DeepEyes* requires neither synthetic reasoning trajectories nor external specialized models to enable this behavior. To guide its reasoning behavior, we propose tool-use-oriented data selection and a reward strategy that promotes effective tool-assisted problem solving. During training, the model's tool use progresses from naive exploration to efficient exploitation, with improved accuracy and visual focus. *DeepEyes* exhibits diverse reasoning behaviors such as visual search and comparison, and achieves competitive results on multiple benchmarks using only a 7B model.

## References


[1] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[2] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[3] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[4] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[5] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

[6] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*, 2024.

[7] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[8] OpenAI. Thinking with images. `https://openai.com/index/thinking-with-images/`, 2025.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[13] Kezhen Chen, Rahul Thapa, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Dragonfly: Multi-resolution zoom supercharges large visual-language model. *arXiv e-prints*, pages arXiv–2406, 2024.

[14] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer, 2024.

[15] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.

[16] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025.

[17] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[19] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[20] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[21] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[22] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[23] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024.

[24] Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*, 2025.

[25] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[26] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[27] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[28] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[29] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[30] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024.

[31] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[32] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *arXiv preprint arXiv:2407.12709*, 2024.

[33] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[34] Chenkai Xu, Xu Wang, Zhenyi Liao, Yishun Li, Tianqi Hou, and Zhijie Deng. Show-o turbo: Towards accelerated unified multimodal understanding and generation. *arXiv preprint arXiv:2502.05415*, 2025.

[35] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[36] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*, 2025.

[37] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[38] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.

[39] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806, 2024.

[40] Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. *arXiv preprint arXiv:2412.02025*, 2024.

[41] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[42] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

[43] Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, et al. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

[44] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. *arXiv preprint arXiv:2504.14920*, 2025.

[45] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18180–18187, 2024.

[46] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. *arXiv preprint arXiv:2412.03548*, 2024.

[47] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

[48] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[49] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[50] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[51] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[52] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

[53] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

[54] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

[55] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[56] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, 2024.

[57] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.

[58] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[59] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025.

[60] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[61] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*, 2024.

[62] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[63] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[64] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[65] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[66] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[67] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[68] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

[69] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[70] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[71] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[72] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[73] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[74] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

# A Prompt

## A.1 System Prompt

SYSTEM_PROMPT

```
You are a helpful assistant.

# Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags:
<tools>
{
  "type": "function",
  "function": {
    "name": "image_zoom_in_tool",
    "description": "Zoom in on a specific region of an image by cropping it
    ↪  based on a bounding box (bbox) and an optional object label.",
    "parameters": {
      "type": "object",
      "properties": {
        "bbox_2d": {
          "type": "array",
          "items": {
            "type": "number"
          },
          "minItems": 4,
          "maxItems": 4,
          "description": "The bounding box of the region to zoom in, as [x1,
          ↪  y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2)
          ↪  is the bottom-right corner."
        },
        "label": {
          "type": "string",
          "description": "The name or label of the object in the specified
          ↪  bounding box (optional)."
        }
      },
      "required": [
        "bbox_2d"
      ]
    }
  }
}
</tools>

# How to call a tool
Return a json object with function name and arguments within
↪  <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

**Example**:
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [10, 20, 100, 200],
↪  "label": "the apple on the desk"}}
</tool_call>
```

### A.2 User Prompt

```
USER_PROMPT
Question: {}

Think first, call **image_zoom_in_tool** if needed, then answer. Format
↪ strictly as: <think>...</think> <tool_call>...</tool_call> (if tools
↪ needed) <answer>...</answer>
```

## B Data Distribution

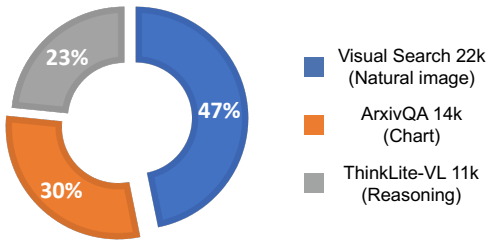



Figure 6: Distribution of training data.

As shown in Figure 6, our training corpus is constructed from three distinct sources, each contributing a unique focus:

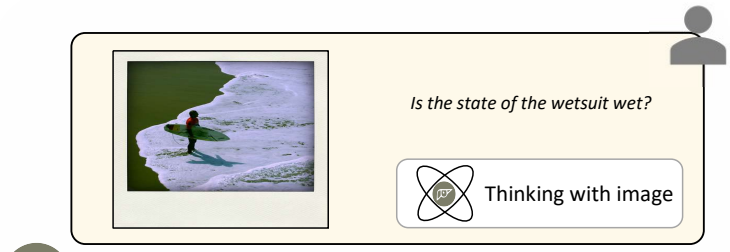
- **Visual Search (47%, 22k samples)**: To support the model's visual grounding and fine-grained perception capabilities, we leverage the $V^*$ dataset [41], which is derived from COCO2017 [74]. This collection emphasizes natural image understanding, where accurate responses require identifying subtle visual cues and object-level distinctions.
- **ArxivQA (30%, 14k samples)**: To diversify the visual input types, we incorporate the ArxivQA dataset [56], which features scientific plots, diagrams, and schematic charts. These samples introduce structured visual semantics beyond natural scenes, enabling the model to better interpret abstract and symbolic visual representations.
- **ThinkLite-VL (23%, 11k samples)**: While the above datasets cover visual understanding and diagram comprehension, they are limited in reasoning variety. To address this, we include multimodal question answering examples from ThinkLite-VL [57], focusing on tasks such as arithmetic reasoning, commonsense inference, and problem solving. This addition is intended to improve general reasoning robustness and mitigate modality-specific overfitting.

## C More Cases

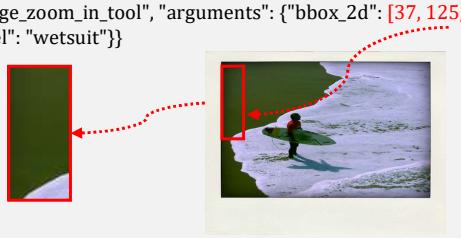
### C.1 Successful Cases

- **Visual Search**

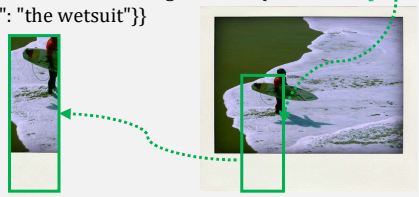
  Figure 7: After an initial observation of the whole image, the model recognized that the current visual information alone was insufficient to determine whether it was wet, and acknowledged that factors such as lighting could cause misleading cues. It was therefore decided that a more detailed inspection was necessary. During the first tool invocation, grounding was inaccurate, and the cropped image failed to provide some clues. The model then conducted a second grounding step, this time actively focusing on the area surrounding the wetsuit in an attempt to locate more direct indicators—such as water droplets or visible signs of wetness. It also incorporated contextual cues from the surrounding environment, such as reflections on wet sand and the wetsuit's contact with water. Ultimately, by combining zoomed-in visual details—such as the wetsuit's dark coloration




<think>
***It's not evident from the image whether the wetsuits are wet.*** Usually, wetsuits that are not wet appear darker, but this could depend on the lighting and the photo itself. Further magnification might be needed to clarify if there are signs of moisture. Calling the zoom-in tool can help explore this aspect more closely.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [37, 125, 81, 231], "label": "wetsuit"}}
</tool_call>

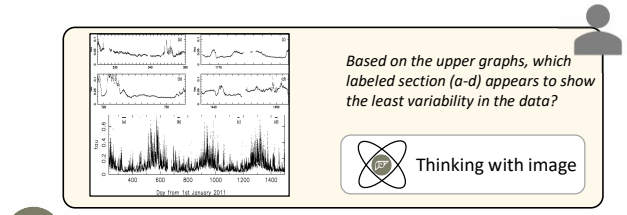


<think>
It seems like the ***image does not provide enough information*** to determine if the wetsuit is wet or not. Since there's a reflection on the wet sand and the wetsuit appears to be in contact with the water, it's plausible, but not guaranteed, that the wetsuit is wet. ***Before confirming whether the wetsuit is wet,*** zooming in on the wetsuit can give more definitive proof or indicators like droplets or moisture on the suit.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [73, 182, 164, 476], "label": "the wetsuit"}}
</tool_call>



<think>
Based on the image, ***the wetsuit appears to be dark and held tightly against the body,*** which could indicate that it is wet. The lighting and the wetness of the sand suggest that the wetsuit might be wet due to recent use near water.
</think>
<answer>
Yes, the state of the wetsuit appears to be wet.
</answer>

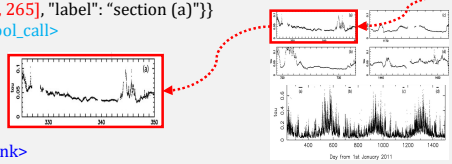
Figure 7: Thinking Pattern: Visual Search.



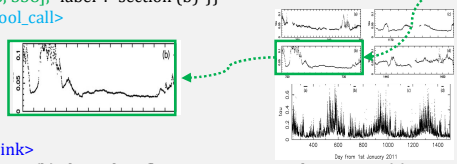
*Based on the upper graphs, which labeled section (a-d) appears to show the least variability in the data?*

Thinking with image

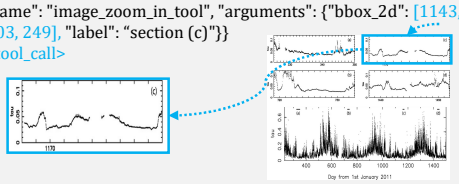
<think>
To determine which section shows the least variability, we need to examine the upper graphs (a, b, c, d). The variability can be assessed by how much the data fluctuates around the mean line. ***Let's observe each section:***
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [124, 47, 955, 265], "label": "section (a)"}}
</tool_call>



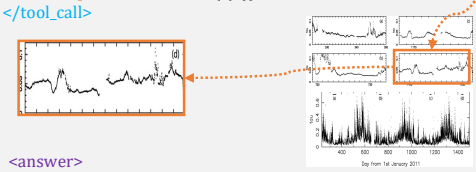
<think>
From the zoomed-in view of section (a), it's clear that the data fluctuates quite a bit, indicating higher variability. Let's proceed to examine the other sections:
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [79, 321, 986, 558], "label": "section (b)"}}
</tool_call>



<think>
Section (b) shows less fluctuation compared to section (a), suggesting less variability. Now, let's zoom in on sections (c) and (d) to compare them more clearly.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [1143, 30, 2003, 249], "label": "section (c)"}}
</tool_call>



<think>
After zooming in on section (c), it appears that the data fluctuates less than in sections (a), (b), and (d), indicating the least variability. Let's ***confirm*** by examining section (d).
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [1090, 317, 2018, 558], "label": "section (a)"}}
</tool_call>



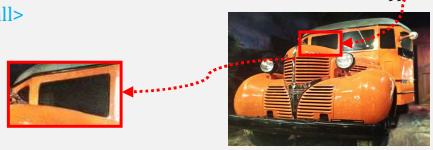
<answer>
Section (c) appears to show the least variability in the data, so the answer is section (c).
</answer>

Figure 8: Thinking Pattern: Visual Comparison.

*Is the window on the front of the vehicle black and square?*

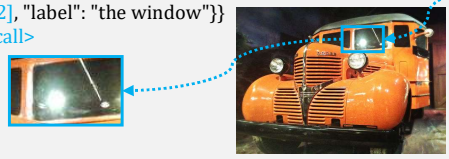
Thinking with image

<think>
Looking at the image, the window on the front of the vehicle appears to be black and square. However, the resolution and angle of the photo may make it ***difficult to see the details clearly***. If we need a closer look, we could use the zoom-in tool to get a better view of the window.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [175, 44, 471,137], "label": "window on the front of the vehicle"}}
</tool_call>



<think>From the zoomed-in image of the vehicle's window, the window appears to be booking a possible diagonal from corner to corner, suggesting at least one dimension ***might not be square***. However, considering the relation of the window to the vehicle's body, which seems to conform to an overall rectangular configuration, the window itself might be square, but confirmation would ***require a closer view or a more specific tool analysis.***
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [222, 51, 455,134], "label": "the window"}}
</tool_call>

<think>The front window on the vehicle ***appears to be*** black and square, judging by the visible parts and typical vehicle design.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [178,37, 342, 112], "label": "the window"}}
</tool_call>



<think>
Upon examining the image more closely, it appears the window on the front of the vehicle is ***indeed black and square.***
</think>
<answer>
Yes, the window on the front of the vehicle is black and square.
</answer>

Figure 9: Thinking Pattern: Visual Confirmation.

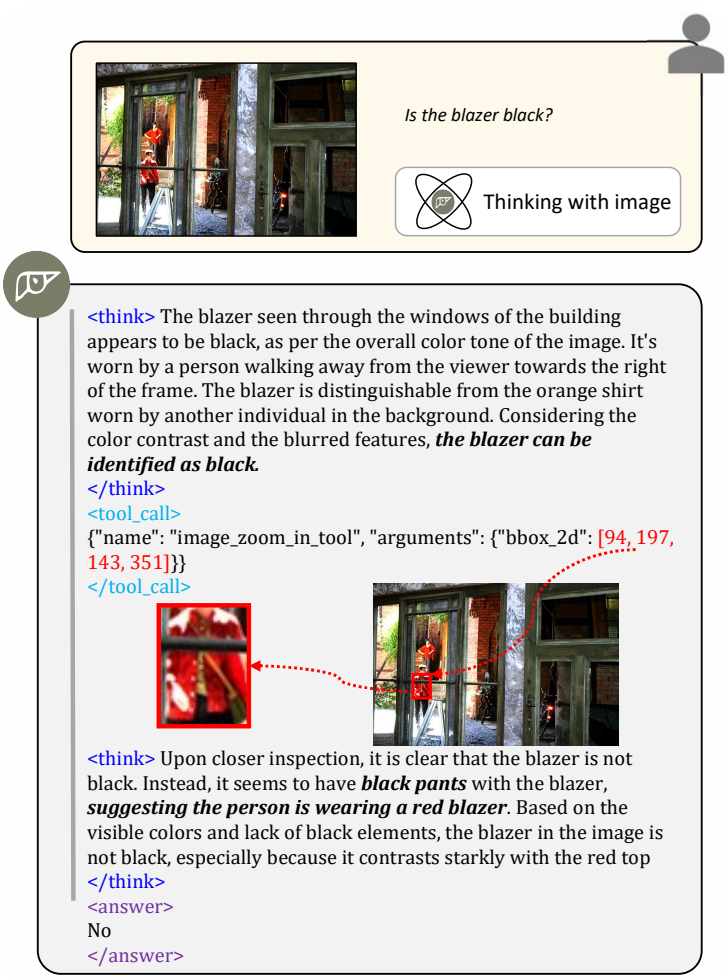



Figure 10: Thinking Pattern: Hallucination Mitigation.

and how it clung to the body—with indirect environmental evidence, the model concluded that the wetsuit appeared to be wet.

- **Visual Comparison**
  Figure 8: To determine which section exhibits the least data variability, the model sequentially zoomed in on the charts of four sections (a, b, c, and d), focusing on fluctuations around the moving average. Through comparison, it found that section (a) showed significant volatility, while section (b) was relatively less volatile. However, section (c) displayed the most stable pattern, with fluctuations clearly smaller than those in the other regions. Based on this analysis, the model concluded that section (c) has the least data variability.
- **Visual Confirmation**
  Figure 9: In this case, the model was initially uncertain about the shape of the window. Through multiple invocations of the zoom-in tool and careful analysis of potential visual details, it gradually resolved its internal uncertainty and ultimately provided a confident answer.
- **Hallucination Mitigation** Figure 10: The model initially confused the colors of the pants and the blazer. However, by leveraging its perceptual capabilities and invoking the zoom-in tool to examine the enlarged region, it ultimately corrected the hallucination.

### C.2 Failed Cases

- **Grounding Limitation**

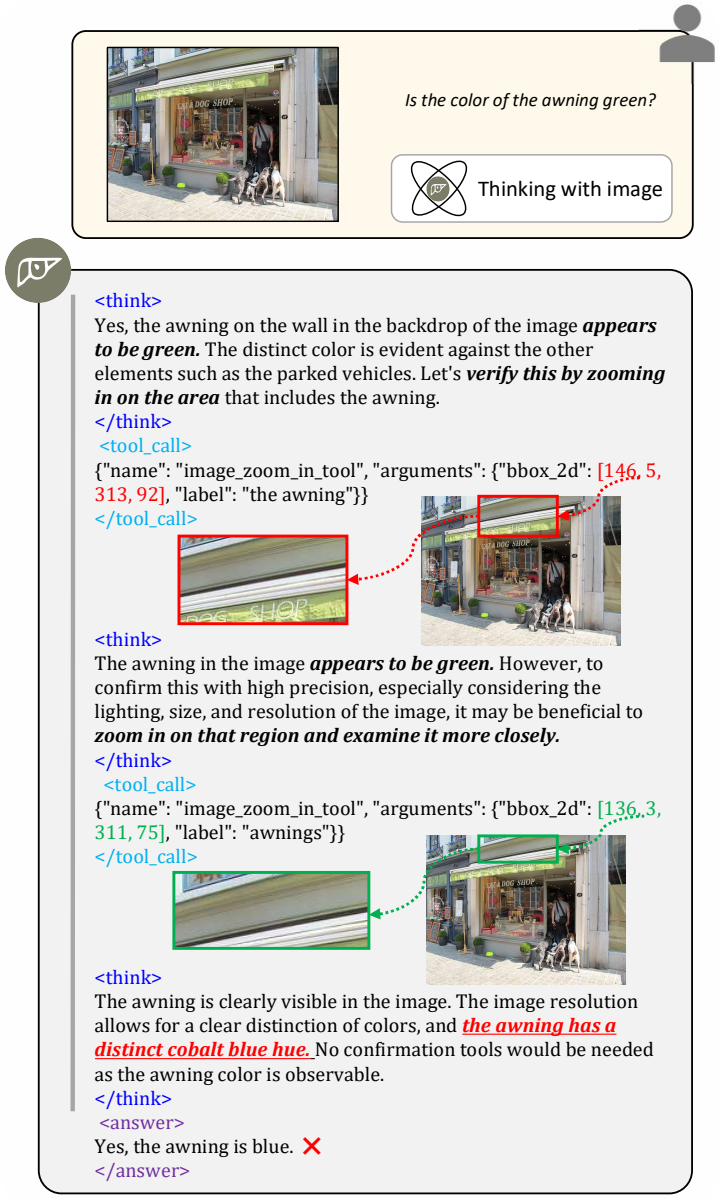
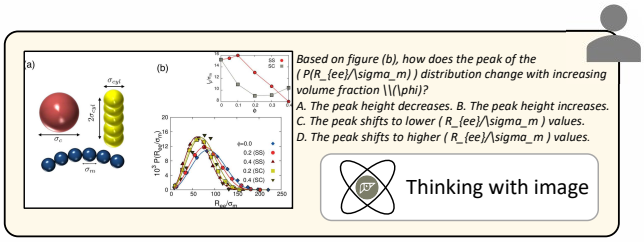
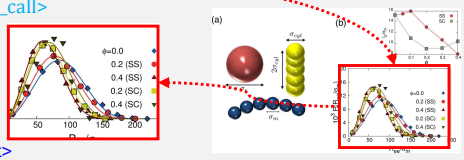


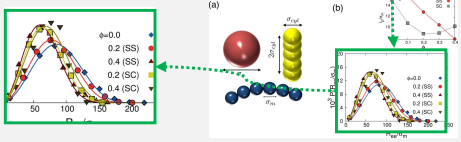

Figure 11: Failure Case: Grounding Limitation.




<think>
Observing figure (b), we need to focus on the position of the peaks of the distributions corresponding to different (\phi). As (\phi) increases from 0.0 to 0.4, the peaks of the distribution shift to higher ( R_{ee}/\sigma_m ) values. This suggests that as the volume fraction (\phi) increases, the most probable ( R_{ee} ) value also increases.Therefore, the correct answer is D.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [1102, 707, 1520, 1057], "label": "peak position"}}
</tool_call>



<think>
The zoomed-in view confirms that as the search space on the horizontal axis for ( R_{ee}/\sigma_m ) increases, the locations of the peaks ***shift to the right.*** This aligns with the observation in the full graph that the peaks move to higher values of ( R_{ee}/\sigma_m ) as (\phi) increases.The peak position for higher values of (\phi) (color-coded differently) shifts to higher ( R_{ee}/\sigma_m ) values compared to lower (\phi).Therefore, the correct answer is D.
</think>
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [1102, 707, 1520, 1057], "label": "peak position"}}
</tool_call>



<answer>
D. The peak shifts to higher ( R_{ee}/\sigma_m ) values. ✗
</answer>

Figure 12: Failure Case: Reasoning Limitation.

Figure 11: The model initially hypothesized that the awning was green. It then invoked the zoom-in tool for a closer inspection, maintaining its assumption while noting the need for more precise verification. However, during the second zoom-in, grounding drift occurred—the awning was no longer within the selected region, and instead, a blue area appeared. This misalignment led to a reversal in the model's judgment, ultimately resulting in an incorrect answer.

- **Reasoning Limitation**
  Figure 12: Although the model was able to accurately locate the position of figure (b) and invoke the tool for detailed inspection, it still lacked fine-grained understanding and reasoning capabilities. It failed to thoroughly analyze the trend changes in the zoomed-in curves, ultimately leading to an incorrect answer.

## D Limitations

Although the simple end-to-end RL can elicit visual reasoning abilities, there still exist shortcuts, such as insufficient richness in the reasoning process and inaccurate target localization. We think these issues stem from limitations in the foundation model's poor capabilities. We only utilized Qwen2.5-VL-7b, which has relatively weak fundamental capabilities due to its small model size.

## E Broader Impacts

Our exploration of interleaved multimodal chain-of-thought reasoning provides valuable insights for the future development of the AI community. By investigating how models can engage in step-by-step visual reasoning through interactive dialogues, we advance understanding of more transparent and interpretable AI systems. This research direction may inspire new architectures and training methodologies that better align with human reasoning processes.

## F Future Work

Currently, our visual reasoning process only includes the crop operation. However, in real-world scenarios, a wider range of tools is needed, such as search and drawing auxiliary lines. We will explore the integration of additional tool utilization in our future work.